WASSERSTEIN ONLINE CHANGE POINT DETECTION WITH NEURAL NETWORKS

Anonymous authors
Paper under double-blind review

ABSTRACT

Online change point detection (CPD) is a fundamental problem in statistics and machine learning. Kernel methods are theoretically principled but are limited by the choice of kernel and lose power in high dimensions. Learning-based approaches, on the other hand, typically lack the same level of theoretical guarantees. To bridge this gap, we propose an online CPD algorithm approximating the Wasserstein distance between two consecutive sliding windows through a Lipschitz-constrained neural network. The network is trained recursively using a small number of gradient iterations per time step, making the detector unsupervised and online. We derive non-asymptotic concentration bounds for the test statistic under the null hypothesis and under the alternative with a partially contaminated window, leveraging a refined analysis with Rademacher complexity bounds. This yields a constant threshold choice that guarantees any prescribed average run length, a logarithmically growing threshold that controls the false alarm probability over an infinite horizon, and an exponential tail bound on the detection delay. These are, to our knowledge, the first results for an online detector whose statistic is computed by a neural network. Simulation results show that the proposed algorithm achieves lower detection delays per average run length than several state-of-the-art detectors.

1 INTRODUCTION

Online change point detection (CPD), also known as quickest detection (Tartakovsky et al., 2014; Xie et al., 2021), monitors a data stream whose statistical distribution may shift abruptly at an unknown time, and aims to raise an alarm as soon as possible after the change occurs while keeping false alarms under control. When the pre- and post-change distributions belong to a known parametric family, classical procedures such as the cumulative sum (CUSUM) statistic (Page, 1954) and the generalized likelihood ratio test (GLRT) (Gustafsson, 1996) are well understood and provably optimal (Lai, 1995; Tartakovsky et al., 2014). In practice, however, the data are often high-dimensional, the change can affect the distribution in an arbitrary way, and no parametric model is available, which calls for non-parametric detectors.

Non-parametric detectors include comparing the empirical distributions of two windows of the stream through a discrepancy between probability measures. The most prominent choice is the kernel maximum mean discrepancy (MMD) (Gretton et al., 2012), which underlies a large family of online detectors with closed-form statistics and well-developed guarantees (Li et al., 2019; Keriven et al., 2020; Wei & Xie, 2026; Kalinke & Gavioli-Akilagun, 2025). The disadvantage is that the quality of the detector is limited by the kernel. Indeed, with a fixed kernel and a bandwidth set by the median heuristic (Garreau et al., 2017), the discrepancy is measured in a data-independent feature space, and its power is known to degrade in high dimensions (Liu et al., 2020; Chang et al., 2019).

This limitation suggests learning the discrepancy from the data instead of fixing it in advance, and neural networks have increasingly been used to this end. Online neural detectors build their statistics from classifiers trained to separate recent from past samples (Hushchyn et al., 2026; Gong et al., 2022; Puchkin & Shcherbakova, 2023), or from neural estimates of density ratios (Wang et al., 2023) or score functions (Zhou et al., 2025). Their theoretical understanding, however, remains limited: existing analyses mostly focus on the offline setting (Li et al., 2024; Kanrar et al., 2025), whereas in the online setting they do not provide control of both detection delay and false alarm rate, and are limited in their choice of divergence. See Appendix A.1 for a detailed review.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Among the learnable discrepancies, the Wasserstein distance (Villani et al., 2008) is a particularly compelling candidate. Beyond its success across statistics and machine learning (Panaretos & Zemel, 2019), it respects the geometry of the sample space, remains finite and continuous between distributions with disjoint supports, and quantifies the magnitude of a shift rather than merely signaling its presence. Moreover, its Kantorovich–Rubinstein dual expresses it as a supremum over 1-Lipschitz functions, and Wasserstein generative adversarial networks (WGANs) (Arjovsky et al., 2017; Gulrajani et al., 2017) show that this supremum can be approximated effectively by a Lipschitz-constrained neural network, so the discrepancy adapts to the data at hand. It is therefore natural to ask whether the Wasserstein distance, learned by a neural network, can serve as the test statistic of an online change point detector, and whether the resulting procedure can be analyzed with the same rigor as its kernel counterparts. We answer both questions affirmatively. A main challenge is that fluctuations under the null depend on how expressive the function class is, and this paper builds upon recent results on Rademacher complexity bounds for neural networks (Golowich et al., 2018) and a refined analysis of their application to the Wasserstein distance to answer these questions. This paper makes the following contributions:

1. **Algorithm development:** We introduce WONDER (**W**asserstein **ON**line change-point **DE**tection with neu**R**al networks), an algorithm in which a Lipschitz-constrained neural network is recursively trained on two consecutive sliding windows to approximate the Wasserstein distance between them, and an alarm is raised when this learned distance exceeds a threshold. The neural network is updated with a few gradient steps per time instant and warm-started from the previous one, so the procedure is fully unsupervised and runs online.
2. **Theoretical analysis:** We establish non-asymptotic concentration bounds for the neural-network Wasserstein statistic that hold with high probability under the null hypothesis, and under the alternative even when the post-change window is only partially contaminated. Building on these bounds, we derive an explicit threshold that guarantees a prescribed ARL, a time-varying threshold that controls the probability of ever raising a false alarm over an infinite horizon, and an exponential tail bound on the detection delay. To the best of our knowledge, these are the first such guarantees for an online detector whose statistic is computed by a neural network.
3. **Empirical validation:** On synthetic streams and on the MNIST data set viewed as a high-dimensional stream, WONDER achieves a lower detection delay than several state-of-the-art detectors at every target ARL, with an advantage that widens as the false alarm requirement becomes more stringent.

2 Problem formulation

We observe an online stream $\mathbf{z}_1, \mathbf{z}_2, \ldots$ of independent random variables belonging to a bounded domain $\mathcal{X} \subset \mathbb{R}^p$. The online CPD problem consists of detecting the existence of a *change point* τ, such that the samples are i.i.d. from a distribution P before τ and i.i.d. from a different distribution Q after τ. Formally, we are concerned with testing *the null hypothesis*:

$$H_0 : \mathbf{z}_t \sim P, \text{ for all } t = 1, 2, \ldots \tag{1}$$

against *the alternative hypothesis*:

$$H_1 : \exists\, \tau \in \mathbb{N}_+, \text{ such that } \mathbf{z}_t \sim \begin{cases} P, & t < \tau, \\ Q, & t \geq \tau. \end{cases} \tag{2}$$

In this setting, the goal is to flag the change point with minimal delay once τ is reached, while avoiding false alarms before the change occurs. Beyond detection, one may also wish to estimate the change point τ as accurately as possible. A decision rule is a stopping time T (i.e., a random time at which the event $\{T \leq t\}$ is determined by the first t observations $\mathbf{z}_1, \ldots, \mathbf{z}_t$ alone). In the sequel, let $\mathbb{P}_\tau$ be the joint distribution of the stream $\{\mathbf{z}_t\}_{t \geq 1}$ when the change point is τ (with $\tau = \infty$ corresponding to the global null), and let $\mathbb{E}_\tau$ be the expectation under this distribution. Following standard sequential testing criteria, T is designed to satisfy one of the following constraints:

1. *Average run length (ARL) control*:

$$\mathbb{E}_\infty[T] \geq \gamma, \qquad \gamma > 1, \tag{3}$$

where $\mathbb{E}_\infty$ is expectation under the global null $\tau = \infty$.

2. *Uniform false-alarm probability (FAP) control*:

$$\mathbb{P}_\infty(T < \infty) \leq \alpha, \qquad \alpha \in (0, 1). \tag{4}$$

A complementary objective is to keep the *detection delay* small by controlling

$$\mathbb{E}_\tau\big[(T - \tau)_+\big]. \tag{5}$$

In this work, we show that the proposed test statistic can be calibrated to satisfy either ARL control or uniform FAP control, while retaining small detection delay with high probability.

3 Proposed Algorithm

At the core of the CPD problem lies a two-sample test between the samples $\mathbf{z}_t$ observed before and after a candidate change point τ'. Since we target non-parametric tests, a natural approach is to compare the two empirical distributions through a distance or divergence.

3.1 Learning Wasserstein distance

In this work, we adopt the Wasserstein distance between two probability measures P, Q supported on $\mathcal{X}$ (Panaretos & Zemel, 2019), which is particularly well suited to non-parametric CPD. It quantifies the minimal cost of transporting mass from one distribution to another. In particular, the Wasserstein-1 distance is defined as (Villani et al., 2008)

$$W(P, Q) = \inf_{\gamma \in \Pi(P,Q)} \mathbb{E}_{(x,y)\sim\gamma}\{\|x - y\|\}\,, \tag{6}$$

where $\Pi(P, Q)$ is the set of all joint distributions γ that have marginals P and Q, respectively.

Unlike the KL (Sugiyama et al., 2007; Liu et al., 2013) or Jensen–Shannon divergences (Puchkin & Shcherbakova, 2023), it remains finite and continuous even when the two distributions have disjoint supports, a situation that is typical for finite samples collected before and after a change. Thus, it quantifies the magnitude of a change rather than merely signaling its presence. However, the infimum in (6) is difficult to compute. Inspired by (Arjovsky et al., 2017), due to the Kantorovich–Rubinstein duality (Villani et al., 2008), the Wasserstein-1 distance can be equivalently expressed as

$$W(P, Q) = \sup_{f \in \mathcal{F}} \Big\{\mathbb{E}_{x\sim P}\{f(x)\} - \mathbb{E}_{y\sim Q}\{f(y)\}\Big\}\,, \tag{7}$$

where $\mathcal{F} = \{\|f\|_L \leq 1\}$ is the set of all real-valued and Lipschitz functions defined on the support $\mathcal{X}$, and $\|f\|_L$ denotes the Lipschitz constant of f, i.e., $\|f\|_L := \sup_{x\neq y} \frac{\|f(x)-f(y)\|}{\|x-y\|}$, and $\|f\|_L \leq 1$ means that f is 1-Lipschitz. The same analysis can be written for a K-Lipschitz case by rescaling (7), one obtains an equivalent formulation up to a multiplicative factor of $1/K$ (Arjovsky et al., 2017).

Remark 3.1 (Connection to the MMD). The dual form (7) is exactly the maximum mean discrepancy (MMD) (Gretton et al., 2012) with the function class $\mathcal{F}$ taken as the unit ball of the Lipschitz seminorm, i.e., the Earth-Mover's distance case discussed in Gretton et al. (2012, Section 7.2.2) and (Sriperumbudur et al., 2010a). The kernel MMD underlying existing nonparametric online CPD methods (Li et al., 2019; Keriven et al., 2020; Kalinke et al., 2025; Kalinke & Gavioli-Akilagun, 2025; Wei & Xie, 2026) instead takes $\mathcal{F}$ to be the unit ball of a reproducing kernel Hilbert space (RKHS). This distinction is key: the 1-Lipschitz ball is a strictly richer, non-smooth class of functions and contains candidates that can concentrate their variation exactly where P and Q differ.

Instead, we consider parameterizing the function $f(\mathbf{x})$ with standard neural networks, of the form

$$f(\mathbf{x}) = \mathbf{W}_d\sigma_{d-1}(\mathbf{W}_{d-1}\sigma_{d-2}(\dots\sigma_1(\mathbf{W}_1\mathbf{x})))\,, \tag{8}$$

where each $\mathbf{W}_i$ is a weight matrix, d is the depth of the network, and each σ_i is a fixed activation function. Restricting the supremum in (7) to the class of 1-Lipschitz neural networks $\mathcal{F}_{\text{NN}} = \{f \text{ as in (8)} : \|f\|_L \leq 1\}$ defines the neural-network Wasserstein distance

$$W_{\text{NN}}(P, Q) = \sup_{f \in \mathcal{F}_{\text{NN}}} \Big\{\mathbb{E}_{x\sim P}\{f(x)\} - \mathbb{E}_{y\sim Q}\{f(y)\}\Big\}\,. \tag{9}$$

Algorithm 1 WONDER: Wasserstein online CPD with neural networks

Input: $\{\mathbf{z}_t\}$, window sizes n, m, critic steps S, step size η, penalty weight λ, thresholds $\{\nu_t\}$.
Initialize critic parameters θ_0.
for $t = n + m, n + m + 1, \ldots$ **do**
  Form consecutive sliding windows $X^{(t)}$ and $Y^{(t)}$ as in (11);
  Set $\theta \leftarrow \theta_{t-1}$;
  for $s = 1, \ldots, S$ **do**
    Gradient ascent on the learning objective: $\theta \leftarrow \theta + \eta \nabla_\theta \mathcal{L}_t(\theta)$ with $\mathcal{L}_t$ in (13);
  Compute the statistic $\hat{g}_t$ using (14);
  if $\hat{g}_t > \nu_t$ **then**
    Flag t as a change point;

Given i.i.d. samples $\mathbf{x}_1, \ldots, \mathbf{x}_n \sim P$ and $\mathbf{y}_1, \ldots, \mathbf{y}_m \sim Q$ associated with the empirical measures $X_{1:n} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ and $Y_{1:m} = \{\mathbf{y}_1, \ldots, \mathbf{y}_m\}$, we estimate $W_{\mathrm{NN}}(P, Q)$ in (9) by its empirical counterpart:

$$\widehat{W}_{\mathrm{NN}}[X_{1:n}|Y_{1:m}] \;=\; \sup_{f \in \mathcal{F}_{\mathrm{NN}}} \left\{ \frac{1}{n} \sum_{i=1}^{n} f(\mathbf{x}_i) - \frac{1}{m} \sum_{j=1}^{m} f(\mathbf{y}_j) \right\}. \tag{10}$$

3.2 Wasserstein online CPD

Our goal is to detect change points by monitoring abrupt changes in the value of $W_{\mathrm{NN}}(P, Q)$ over time. Concretely, we use two consecutive sliding windows of fixed lengths n and m: for each time step $t \geq n + m$, define

$$X^{(t)} = \{\mathbf{z}_{t-n-m+1}, \ldots, \mathbf{z}_{t-m}\} \qquad \text{and} \qquad Y^{(t)} = \{\mathbf{z}_{t-m+1}, \ldots, \mathbf{z}_t\}. \tag{11}$$

We quantify their discrepancy via the empirical neural-network Wasserstein statistic (10) applied to the two windows:

$$g_t = \widehat{W}_{\mathrm{NN}}\big[X^{(t)}|Y^{(t)}\big] := \sup_{f \in \mathcal{F}_{\mathrm{NN}}} \left\{ \frac{1}{n} \sum_{i=t-n-m+1}^{t-m} f(\mathbf{z}_i) - \frac{1}{m} \sum_{j=t-m+1}^{t} f(\mathbf{z}_j) \right\}. \tag{12}$$

We note that computing the exact Lipschitz constant of a neural network is NP-hard in general Virmaux & Scaman (2018), but we can enforce a Lipschitz constraint by imposing appropriate constraints on the weight matrices and activation functions, as commonly done in the literature of Wasserstein GANs (Arjovsky et al., 2017; Gulrajani et al., 2017). In this work, we consider enforcing the 1-Lipschitz constraint using gradient penalty (Gulrajani et al., 2017).

Specifically, following the two-sided penalty of (Gulrajani et al., 2017), we augment the dual objective with a term that drives the gradient norm of the critic towards 1 on segments joining the two windows: let $\hat{\mathbf{z}} = u\,\mathbf{x} + (1-u)\,\mathbf{y}$, where $u \sim \mathrm{Uniform}[0, 1]$ and $\mathbf{x}$, $\mathbf{y}$ are drawn uniformly from $X^{(t)}$ and $Y^{(t)}$, respectively. The learning objective becomes

$$\mathcal{L}_t(\theta) = \frac{1}{n} \sum_{\mathbf{x} \in X^{(t)}} f_\theta(\mathbf{x}) - \frac{1}{m} \sum_{\mathbf{y} \in Y^{(t)}} f_\theta(\mathbf{y}) - \lambda\, \mathbb{E}_{\hat{\mathbf{z}}} \Big[\big(\|\nabla_{\hat{\mathbf{z}}} f_\theta(\hat{\mathbf{z}})\| - 1 \big)^2 \Big], \tag{13}$$

where $\lambda > 0$ is the penalty weight and the expectation is approximated by a fixed number of random interpolates per update. The penalty keeps f_θ approximately 1-Lipschitz.

After S optimization steps, with parameter $\hat{\theta}_t$, we evaluate the statistic without the penalty term:

$$\hat{g}_t = \frac{1}{n} \sum_{\mathbf{x} \in X^{(t)}} f_{\hat{\theta}_t}(\mathbf{x}) - \frac{1}{m} \sum_{\mathbf{y} \in Y^{(t)}} f_{\hat{\theta}_t}(\mathbf{y}), \tag{14}$$

and raise an alarm when $\hat{g}_t > \nu_t$. The complete CPD procedure, which we refer to as WONDER, is outlined in Algorithm 1.

4 THEORETICAL ANALYSIS

In this section, we assess the performance of the proposed the test statistics g_t in two main aspects: i) the average run length (ARL) and false alarm probability (FAP) under the null hypothesis, and ii) the detection delay under the alternative hypothesis. To achieve this, we first need two supplementary outcomes: non-asymptotic concentration bounds for the test ststistics g_t, under both iii) the null hypothesis and iv) under the alternative. All proofs are deferred to Appendix A.2.

4.1 ASSUMPTIONS AND AUXILIARY RESULTS

First, we make the following mild assumptions on the activation functions, the weight matrices, the Lipschitzness of neural networks, and the data domain. Then, we define the *Rademacher complexity*, which will be used to control the complexity of the function class $\mathcal{F}_{\text{NN}}$.

Assumption 4.1 (Network)**.** (Golowich et al., 2018) For every layer ℓ, the activation function σ_ℓ, satisfying $\sigma_\ell(\mathbf{0}) = \mathbf{0}$, is element-wise and Lipschitz continuous with a Lipschitz constant of at most 1, and positive-homogeneous, i.e., $\sigma_\ell(\alpha z) = \alpha\sigma_\ell(z)$ for all $\alpha \geq 0$ and $z \in \mathbb{R}$. Meanwhile, the weight matrix $\mathbf{W}_\ell$ satisfies $\|\mathbf{W}_\ell\|_F \leq M_\ell$ for a given constant $M_\ell > 0$.

An important example of the above is ReLU networks, where every σ_ℓ corresponds to applying the (positive-homogeneous) ReLU function $\sigma(z) = \max\{0, z\}$ on each element.

Assumption 4.2 (Lipschitzness)**.** (Arjovsky et al., 2017) Every network f of the form (8) in the class $\mathcal{F}_{\text{NN}}$ is 1-Lipschitz, i.e., $\|f\|_L \leq 1$, so that for all $\mathbf{x}, \mathbf{y} \in \mathcal{X}$, we have $|f(\mathbf{x}) - f(\mathbf{y})| \leq \|\mathbf{x} - \mathbf{y}\|$.

To enforce a Lipschitz constraint of $\mathbf{W}_\ell$ in practice, several approaches can be used, ranging from simplistic clipping of the weight matrix after each gradient update as in (Arjovsky et al., 2017) to more sophisticated methods such as gradient penalty (Gulrajani et al., 2017).

Assumption 4.3 (Bounded domain)**.** (Golowich et al., 2018) There exists a constant $B > 0$ such that the data domain $\mathcal{X} \subset \mathbb{R}^p$ is bounded and satisfies

$$\mathcal{X} \subseteq \{\mathbf{x} \in \mathbb{R}^p : \|\mathbf{x}\| \leq B\}. \tag{15}$$

This assumption holds whenever the observations are physically bounded, both before and after the change, e.g., pixel intensities or sensor readings with a finite dynamic range.

Rademacher complexity: The empirical Rademacher complexity of a function class $\mathcal{F}$ with respect to a sample $X_{1:n} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ with size n distributed i.i.d. according to some measure U is defined as (Bartlett & Mendelson, 2002; Koltchinskii, 2002)

$$\hat{\mathcal{R}}_n(\mathcal{F}) = \mathbb{E}_{\boldsymbol{\sigma}}\left\{ \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \sigma_i f(\mathbf{x}_i) \right\}, \tag{16}$$

where $\boldsymbol{\sigma} = \{\sigma_1, \ldots, \sigma_n\}$ are independent Rademacher random variables uniformly chosen from $\{-1, 1\}$, that is, $\mathbb{P}(\sigma_i = 1) = \mathbb{P}(\sigma_i = -1) = 1/2$. The Rademacher complexity of $\mathcal{F}$ is defined as

$$\mathcal{R}_{U,n}(\mathcal{F}) = \mathbb{E}_U\{\hat{\mathcal{R}}_n(\mathcal{F})\}. \tag{17}$$

Note that we include an explicit definition of the measure to which the Rademacher complexity is being computed in the notation $\mathcal{R}_{U,n}$. This is because the CPD problem deals with two different distributions, thus, specifying which data distribution is being considered becomes necessary.

Under Assumptions 4.1 and 4.3, we are ready to introduce a bound on the Rademacher complexity of the function class $\mathcal{F}_{\text{NN}}$ induced by the neural network parametrization in (8). The following proposition is a specialization to our settings of the general bound from Theorem 1 in (Golowich et al., 2018).

Proposition 4.4. *Let $\mathcal{F}_{\text{NN}}$ be the function class induced by the neural network parametrization in* (8) *with data points $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathcal{X}$. Under Assumptions 4.1 and 4.3, the empirical Rademacher complexity of $\mathcal{F}_{\text{NN}}$ satisfies:*

$$\hat{\mathcal{R}}_n(\mathcal{F}_{\text{NN}}) \leq \frac{B \prod_{\ell=1}^{d} M_\ell \left(\sqrt{2\log(2)d} + 1\right)}{\sqrt{n}}. \tag{18}$$

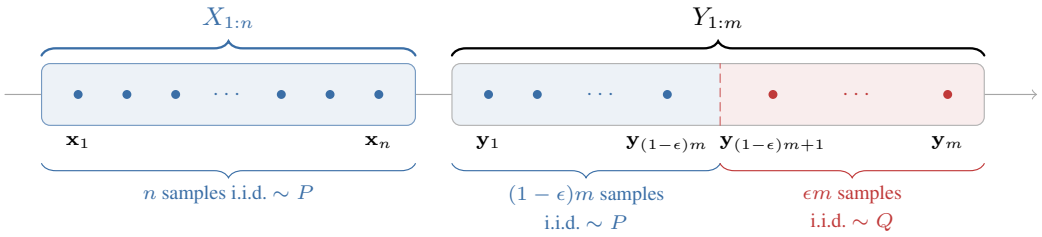


Figure 1: Illustration of the contaminated two-sample setting: the samples in $X_{1:n}$ are drawn i.i.d. from P, while $Y_{1:m}$ contains $(1-\epsilon)m$ samples drawn i.i.d. from P followed by ϵm samples drawn i.i.d. from Q, where the degree of contamination $\epsilon \in (0, 1]$ is such that ϵm is an integer.

Since the bound in (18) holds uniformly over the data points in $\mathcal{X}$, it also upper bounds the population Rademacher complexity $\mathcal{R}_{P,n}(\mathcal{F}_{\mathrm{NN}})$ for any distribution P supported on $\mathcal{X}$, a fact we will use repeatedly in the sequel.

4.2 Concentration bounds

A key tool for our theoretical analysis is concentration results for the Wasserstein distance. While such problems have been studied in the case of optimal transport (Bolley et al., 2007; Boissard, 2011; Sriperumbudur et al., 2009), here we study the dual form of the problem in which the function class is replaced by feedforward neural networks.

Under the null hypothesis: We first consider the null hypothesis, where both windows are drawn from the same distribution, and then the alternative hypothesis, where the second window is contaminated by post-change samples. The following theorem gives a high-probability upper bound on $\widehat{W}_{\mathrm{NN}}[X_{1:n}|Y_{1:m}]$ in (10).

Theorem 4.5. *Let $X_{1:n}$ and $Y_{1:m}$ be two independent samples, each consisting of i.i.d. entries drawn from the same distribution P on $\mathcal{X}$. Under Assumptions 4.1–4.3, for any $\delta \in (0, 1)$, with probability at least $1 - \delta$,*

$$\widehat{W}_{\mathrm{NN}}[X_{1:n}|Y_{1:m}] \;\le\; 2\Big(\hat{\mathcal{R}}_n(\mathcal{F}_{\mathrm{NN}}) + \hat{\mathcal{R}}_m(\mathcal{F}_{\mathrm{NN}})\Big) + 3\sqrt{2}\,B\Big(\frac{1}{\sqrt{n}} + \frac{1}{\sqrt{m}}\Big)\sqrt{\ln\Big(\frac{3}{\delta}\Big)}. \tag{19}$$

Theorem 4.5 bounds $\widehat{W}_{\mathrm{NN}}[X_{1:n}|Y_{1:m}]$ in terms of the empirical Rademacher complexities of the neural network class $\mathcal{F}_{\mathrm{NN}}$. To obtain an explicit concentration inequality, we replace these complexities using the deterministic distribution-free estimate in Proposition 4.4. This gives:

Corollary 4.6. *Let $X_{1:n}$ and $Y_{1:m}$ be two independent samples, each consisting of i.i.d. entries drawn from the same distribution P on $\mathcal{X}$. Under Assumptions 4.1–4.3, for any $\delta \in (0, 1)$, with probability at least $1 - \delta$,*

$$\widehat{W}_{\mathrm{NN}}[X_{1:n}|Y_{1:m}] \;\le\; B\Big(\frac{1}{\sqrt{n}} + \frac{1}{\sqrt{m}}\Big)\left[2\prod_{\ell=1}^{d} M_\ell\big(\sqrt{2\log(2)\,d}+1\big) + 3\sqrt{2\,\ln\Big(\frac{3}{\delta}\Big)}\right]. \tag{20}$$

Under the alternative hypothesis: We next turn to the alternative hypothesis, and derive a high-probability lower bound on $\widehat{W}_{\mathrm{NN}}[X_{1:n}|Y_{1:m}]$ in (10) when the second window is contaminated by samples from $Q \neq P$, as visualized in Figure 1.

Theorem 4.7. *Let P and Q be two probability distributions supported on $\mathcal{X}$, and let $X_{1:n} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ and $Y_{1:m} = \{\mathbf{y}_1, \ldots, \mathbf{y}_m\}$ be two sets of independent samples distributed as follows:*

- *the samples in $X_{1:n}$ are drawn i.i.d. from distribution P.*
- *the samples in $Y_{1:m}$ are drawn as: $\mathbf{y}_1, \ldots, \mathbf{y}_{(1-\epsilon)m}$ be drawn i.i.d. from distribution P, while $\mathbf{y}_{(1-\epsilon)m+1}, \ldots, \mathbf{y}_m$ be drawn i.i.d. from distribution Q.*

where the "degree of contamination" $\epsilon \in (0, 1]$ such that ϵm is an integer. Under Assumptions 4.1–4.3, for any $\delta \in (0, 1)$, with probability at least $1 - \delta$:

$$\widehat{W}_{\mathrm{NN}}[X_{1:n}|Y_{1:m}] \geq \epsilon\, W_{\mathrm{NN}}(P, Q) - 2\mathcal{R}_{P,n}(\mathcal{F}_{\mathrm{NN}}) - 2\mathcal{R}_{Y,m}(\mathcal{F}_{\mathrm{NN}}) - \sqrt{\frac{c_{xy}^2\, \ln(1/\delta)}{2}}, \tag{21}$$

where $c_{xy}^2 := 4B^2\big(\frac{1}{n} + \frac{1}{m}\big)$, and $\mathcal{R}_{Y,m}(\mathcal{F}_{\mathrm{NN}})$ denotes the Rademacher complexity of the contaminated second window, i.e., computed with the expectation taken over the product law $P^{\otimes(1-\epsilon)m} \otimes Q^{\otimes \epsilon m}$ of the samples $Y_{1:m}$.

Combining this result with the distribution-free upper bound on the Rademacher complexity in Proposition 4.4, we obtain the following corollary.

Corollary 4.8. *Let P, Q, $X_{1:n}$, and $Y_{1:m}$ be as in Theorem 4.7. Under Assumptions 4.1–4.3, for any $\delta \in (0, 1)$, with probability at least $1 - \delta$,*

$$\begin{aligned}\widehat{W}_{\mathrm{NN}}[X_{1:n}\,|\,Y_{1:m}] \;\geq\; & \epsilon\, W_{\mathrm{NN}}(P, Q) - 2B \prod_{\ell=1}^{d} M_\ell\big(\sqrt{2\log(2)\, d} + 1\big) \left(\frac{1}{\sqrt{n}} + \frac{1}{\sqrt{m}}\right) \\ & - \sqrt{2}\, B \sqrt{\Big(\frac{1}{n} + \frac{1}{m}\Big) \ln\Big(\frac{1}{\delta}\Big)}\,. \end{aligned} \tag{22}$$

4.3 Theoretical Results of CPD

Leveraging the previous results, we now present results on the ARL, FAP and detection delay that quantify the performance bounds of the test statistics g_t. In the remainder of this section, we shift the time origin so that the tests are indexed by $t = 1, 2, \ldots$, with $t = 1$ corresponding to the first time step, $n + m$ in the original time axis, at which both windows in (11) are available.

Under the null hypothesis: We first work under the global null hypothesis H_0, where the entire stream is i.i.d. from a fixed distribution P, and quantify the false-alarm behavior of the detector through its ARL and FAP. The stopping time, defined as

$$T := \inf\{t \geq 1 :\; g_t > \nu_t\}\,, \tag{23}$$

declares a change whenever g_t exceeds a threshold $\nu_t > 0$. We start with the ARL $\mathbb{E}_\infty[T]$, i.e., the expected number of tests performed before a false alarm is raised. The following result shows that a constant threshold ν can be chosen to guarantee any prescribed ARL level γ.

Theorem 4.9 (Average run length). *Let T be the stopping time* (23) *with a constant threshold $\nu_t \equiv \nu$, and let $\gamma \geq \frac{1}{2}$. Under Assumptions 4.1–4.3, if ν is chosen such that $\nu \geq C_{n,m}\big(C_d + 3\sqrt{2}\,\sqrt{\ln(6\gamma)}\big)$, where $C_{n,m} := B\big(\frac{1}{\sqrt{n}} + \frac{1}{\sqrt{m}}\big)$ and $C_d := 2\prod_{\ell=1}^{d} M_\ell\big(\sqrt{2\log(2)\, d} + 1\big)$, then $\mathbb{E}_\infty[T] \geq \gamma$.*

Still under H_0, we now control the FAP, i.e., the probability of *ever* raising a false alarm over an infinite horizon: $\mathbb{P}_\infty(T < \infty) \leq \alpha$ for $\alpha \in (0, 1)$. The following result shows that this is achieved by a time-varying threshold ν_t that grows only logarithmically in t.

Theorem 4.10 (False alarm probability). *Let $\alpha \in (0, 1)$, and let T be the stopping time* (23) *with the time-varying thresholds $\nu_t := C_{n,m}\big(C_d + 3\sqrt{2}\,\sqrt{\ln(3t(t+1)/\alpha)}\big)$. Under Assumptions 4.1–4.3, T satisfies the infinite-horizon false-alarm guarantee $\mathbb{P}_\infty(T < \infty) \leq \alpha$.*

Under the alternative hypothesis: We now turn to the alternative hypothesis H_1, and study the detection delay incurred by the stopping time T in (23) when the threshold sequence ν_t is chosen as in Theorem 4.10 to control the uniform FAP at some level $\alpha \in (0, 1)$, with consecutive windows defined in (11) as illustrated in Figure 4 in Appendix A.2. Index the post-change window positions by $t_s := t_0 + s - 1$ for $s \geq 1$, so that the right window $Y^{(t_s)}$ contains a fraction $\epsilon_{t_s} := s/m$ of post-change samples for $1 \leq s \leq m$, and let $D := \inf\{s \geq 1 :\; g_{t_s} > \nu_{t_s}\} - 1$ be the detection delay, i.e., the number of post-change window positions at which no alarm is raised. The following theorem provides an exponential tail bound on D.

Theorem 4.11 (Detection delay). *For an integer $1 \le k \le m$, define the margin $\Delta_{t_k} := \epsilon_{t_k} W_{\mathrm{NN}}(P,Q) - C_{n,m} C_d - \nu_{t_k}$, and let c_{xy} be as in Theorem 4.7. Under Assumptions 4.1–4.3,*

$$\mathbb{P}(D \ge k) \le \exp\Big(- \frac{2(\Delta_{t_k})_+^2}{c_{xy}^2} \Big), \tag{24}$$

where $(u)_+ := \max\{u, 0\}$.

Remark 4.12 (Non-triviality of the bound). Since $(\Delta_{t_k})_+ \ge 0$, the right-hand side of (24) never exceeds 1, so the bound is always a valid probability bound; it is strictly smaller than 1 whenever $\Delta_{t_k} > 0$, i.e., $\epsilon_{t_k} W_{\mathrm{NN}}(P,Q) > C_{n,m}\big(2C_d + 3\sqrt{2}\sqrt{\ln(3t_k(t_k+1)/\alpha)}\big)$, which holds for sufficiently large windows since $C_{n,m} = B\big(\frac{1}{\sqrt{n}} + \frac{1}{\sqrt{m}}\big)$.

5 Experiments

In this section, we evaluate WONDER (Algorithm 1) on synthetic data streams and on the MNIST data set. We compare against five state-of-the-art nonparametric online CPD methods: four kernel-based detectors, namely online RFF-MMD (Kalinke & Gavioli-Akilagun, 2025), the Scan B-statistic (ScanB) (Li et al., 2019), online kernel CUSUM (OKCUSUM) (Wei & Xie, 2026), and NEWMA (Keriven et al., 2020), and the neural-network-based NN-CUSUM (Gong et al., 2022). The kernel-based baselines use the Gaussian kernel $K(\mathbf{x}, \mathbf{y}) = e^{-\gamma \|\mathbf{x}-\mathbf{y}\|_2^2}$, or its random Fourier feature approximation, with the bandwidth γ set by the median heuristic Garreau et al. (2017).

Evaluation protocol. Following Kalinke & Gavioli-Akilagun (2025); Wei & Xie (2026), we assess each method through the trade-off between the average run length (ARL) under the null and the expected detection delay (EDD) under the alternative. For a fair comparison, we calibrate the thresholds of all methods by Monte Carlo simulations. Specifically, for each algorithm we process 100 independent null streams of $2^{10} - 1$ samples, each preceded by a warm-up sequence of $1{,}000$ samples (used as the reference sample by ScanB and OKCUSUM), and set the threshold associated with a target ARL γ to the empirical $(1 - 1/\gamma)$-quantile of the pooled test statistics. To estimate the EDD, we then process 64 pre-change samples from P followed by 64 post-change samples from Q, and record the index at which the statistic first exceeds the threshold, with results averaged over 100 independent runs. NEWMA processes 400 additional pre-change samples in both experiments to reduce the variance of its initial statistics. For NN-CUSUM, we use the first 500 warm-up samples to form its reference pool, and its classifier is trained on the remaining 500 warm-up samples.

5.1 Synthetic data

Setup. We reproduce the synthetic setting of (Wei & Xie, 2026; Kalinke & Gavioli-Akilagun, 2025): the pre-change distribution is $P = N(\mathbf{0}, \mathbf{I}_{20})$ and, with $\sigma = 2$ and $p = 20$, the post-change distribution Q is one of (i) the Gaussian mixture $0.7\, N(\mathbf{0}, \sigma^2 \mathbf{I}_{20}) + 0.3\, N(\mathbf{0}, \mathbf{I}_{20})$, (ii) $\mathrm{Laplace}(0, \sigma)^p$ with independent coordinates, or (iii) $\mathrm{Uniform}(-\sigma^2, \sigma^2)^p$ with independent coordinates. The kernel-based baselines use the parameters recommended in (Kalinke & Gavioli-Akilagun, 2025). We select the parameters of NN-CUSUM by a grid search to obtain the best performance. For WONDER, the critic f_θ is a bias-free ReLU multilayer perceptron with hidden layer sizes $(256, 128, 64)$, matching the parametrization in (8), initialized with orthogonal weights. We use window sizes $n = 32$ and $m = 2$ (see the ablation study in Appendix A.3.1), and at each time step perform $S = 8$ SGD steps with step size $\eta = 10^{-2}$ on the penalized objective (13) with gradient penalty weight $\lambda = 2$.

Results. Figure 2 reports the EDD of each method as a function of the target ARL. WONDER achieves the smallest detection delay among all methods for every alternative and essentially every target ARL, the only exception being the Uniform alternative at the two smallest target ARLs, where RFF-MMD is marginally faster. For the Laplace and Uniform alternatives, it detects the change almost instantaneously across the entire ARL range. The largest gains are observed for the more challenging mixed-normal alternative, where WONDER detects the change noticeably faster than all baselines at large target ARL. Moreover, the EDD of WONDER grows very slowly with the target ARL, indicating a clear separation between the null and post-change distributions of the test statistic, consistent with the concentration results of Section 4.

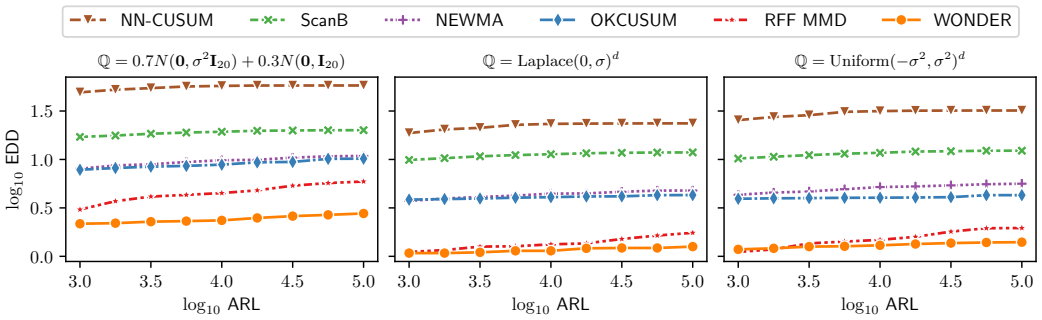


Figure 2: EDD versus target ARL from $P = N(\mathbf{0}, \mathbf{I}_{20})$ to Q indicated on top ($d = 20$, $\sigma = 2$).

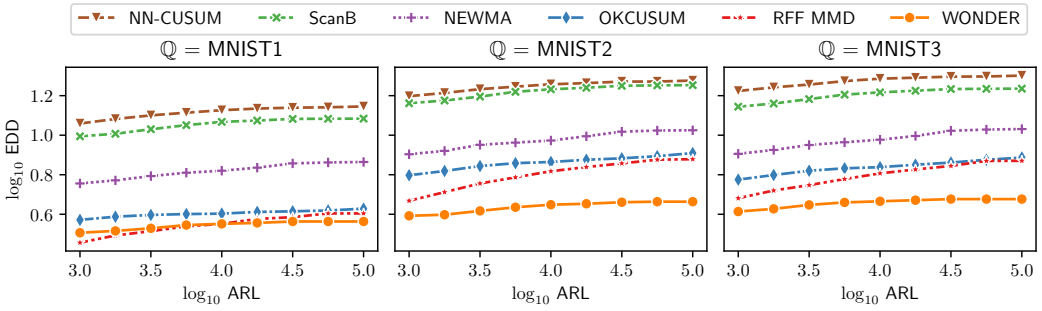


Figure 3: EDD versus target ARL from digit 0 to digits 1, 2, and 3 (left to right).

5.2 MNIST DATA

Setup. Following Wei & Xie (2026); Kalinke & Gavioli-Akilagun (2025), we interpret the MNIST data set (LeCun, 1998) as a high-dimensional data stream ($p = 784$): the stream consists of images of the digit 0 (pre-change), and after the change point the images depict a different digit $q \in \{1, \ldots, 9\}$ (post-change). The kernel-based baselines use the parameters as in (Kalinke & Gavioli-Akilagun, 2025), and the parameters of NN-CUSUM are selected by a grid search. For WONDER, we use a critic with hidden layer sizes $(512, 128)$, window sizes $n = 256$ and $m = 5$, $S = 8$ SGD steps per time step with step size $\eta = 7 \times 10^{-3}$, and gradient penalty weight $\lambda = 0.15$.

Results. Figure 3 reports the EDD as a function of the target ARL for post-change digits 1–3; the remaining digits 4–9 show the same trends and are deferred to Appendix A.3. WONDER achieves the lowest EDD on all nine post-change digits over essentially the whole ARL range, with the single exception of digit 1 at small ARL, where RFF-MMD is marginally faster. The improvements are most pronounced on the other digits. As in the synthetic experiments, the EDD curves of WONDER are nearly flat in the target ARL, so the advantage widens as the false-alarm requirement becomes more stringent.

6 DISCUSSION

Conclusion. We introduced WONDER, an online CPD algorithm whose statistic is the Wasserstein distance between two sliding windows, learned by a Lipschitz-constrained neural network. We proved non-asymptotic concentration bounds for this statistic under the null and alternative hypotheses, in which the network enters only through its depth and the product of the Frobenius norms of its weights, and derived from them explicit thresholds controlling the ARL and the FAP, together with an exponential tail bound on the detection delay. On synthetic streams and on MNIST, WONDER achieves lower delays than baselines, with nearly flat delay curves in the target ARL.

Limitations. The analysis concerns the exact supremum g_t in (12), whereas Algorithm 1 computes $\hat{g}_t$ in (14) after S gradient descent steps. Bounding this optimization gap is left for future work.

AI USE STATEMENT

(This section is **required** and does not count toward the page limit.)

In this work, we used generative AI tools to assist in implementing the methods and to improve the readability of the paper. We have not used generative AI tools for any other task with required disclosure. We have reviewed all AI-assisted work: the code was tested by the authors, and all text edits were checked to preserve the technical content. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

(This section is **recommended** and does not count toward the page limit.)

This paper presents work that aims to advance the field of machine learning, especially in change point detection. There are many potential societal consequences of our work, none of which we feel must be specifically highlighted here.

REPRODUCIBILITY STATEMENT

(This section is **recommended** and does not count toward the page limit.)

The proposed method is fully specified in Algorithm 1 (Section 3). All assumptions underlying the theoretical results are stated in Section 4.1, and complete proofs are given in Appendix A.2. The evaluation protocol, data sets, network architectures, and all hyperparameters are described in Section 5; the experiments use only synthetic data and the public MNIST data set. The source code will be released upon acceptance.

REFERENCES

Nauman Ahad, Eva L Dyer, Keith B Hengen, Yao Xie, and Mark A Davenport. Learning sinkhorn divergences for change point detection. *IEEE Transactions on Signal Processing*, 2025.

Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *International conference on machine learning*, pp. 214–223. PMLR, 2017.

Sylvain Arlot, Alain Celisse, and Zaid Harchaoui. A kernel multiple change-point algorithm via model selection. *Journal of machine learning research*, 20(162):1–56, 2019.

Peter L Bartlett and Shahar Mendelson. Rademacher and Gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3(Nov):463–482, 2002.

Emmanuel Boissard. Simple bounds for the convergence of empirical and occupation measures in 1-wasserstein distance. *Electronic Journal of Probability*, 16:2296–2333, 2011.

François Bolley, Arnaud Guillin, and Cédric Villani. Quantitative concentration inequalities for empirical measures on non-compact spaces. *Probability Theory and Related Fields*, 137(3): 541–593, 2007.

Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration Inequalities*. Oxford University Press, 2013.

Wei-Cheng Chang, Chun-Liang Li, Yiming Yang, and Barnabás Póczos. Kernel change-point detection with auxiliary deep generative models. In *International Conference on Learning Representations*, 2019.

Hao Chen. Sequential change-point detection based on nearest neighbors. *The Annals of Statistics*, 47(3):1381–1407, 2019.

Xiuyuan Cheng and Alexander Cloninger. Classification logit two-sample testing by neural networks for differentiating near manifold densities. *IEEE Transactions on Information Theory*, 68(10): 6631–6662, 2022.

AFB Costa and MA Rahim. A single EWMA chart for monitoring process mean and process variance. *Quality Technology & Quantitative Management*, 3(3):295–305, 2006.

Eustasio Del Barrio, Evarist Giné, and Carlos Matrán. Central limit theorems for the wasserstein distance between the empirical and the true distributions. *Annals of Probability*, pp. 1009–1071, 1999.

Kamil Faber, Roberto Corizzo, Bartlomiej Sniezynski, Michael Baron, and Nathalie Japkowicz. Watch: Wasserstein change point detection for high-dimensional time series data. In *2021 IEEE international conference on big data (Big data)*, pp. 4450–4459. IEEE, 2021.

André Ferrari, Cédric Richard, and Louis Verduci. Distributed change detection in streaming graph signals. In *2019 IEEE 8th International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP)*, pp. 166–170. IEEE, 2019.

Damien Garreau, Wittawat Jitkrittum, and Motonobu Kanagawa. Large sample analysis of the median heuristic. *arXiv preprint arXiv:1707.07269*, 2017.

Aude Genevay, Lénaic Chizat, Francis Bach, Marco Cuturi, and Gabriel Peyré. Sample complexity of sinkhorn divergences. In *The 22nd international conference on artificial intelligence and statistics*, pp. 1574–1583. PMLR, 2019.

Jialiang Geng and George Michailidis. Neural network-based change point detection for large-scale time-evolving data. *arXiv preprint arXiv:2503.09541*, 2025.

Noah Golowich, Alexander Rakhlin, and Ohad Shamir. Size-independent sample complexity of neural networks. In *Conference On Learning Theory*, pp. 297–299. PMLR, 2018.

Dian Gong, Gérard Medioni, Sikai Zhu, and Xuemei Zhao. Kernelized temporal cut for online temporal segmentation and recognition. In *Computer Vision–ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part III 12*, pp. 229–243. Springer, 2012.

Tingnan Gong, Junghwan Lee, Xiuyuan Cheng, and Yao Xie. Neural network-based cusum for online change-point detection. *arXiv preprint arXiv:2210.17312*, 2022.

Arthur Gretton, Karsten Borgwardt, Malte Rasch, Bernhard Schölkopf, and Alex Smola. A kernel method for the two-sample-problem. *Advances in neural information processing systems*, 19, 2006.

Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *The journal of machine learning research*, 13:723–773, 2012.

Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville. Improved training of Wasserstein GANs. *Advances in neural information processing systems*, 30, 2017.

Fredrik Gustafsson. The marginalized likelihood ratio test for detecting abrupt changes. *IEEE Transactions on automatic control*, 41(1):66–78, 1996.

Fredrik Gustafsson. *Adaptive filtering and change detection*, volume 1. Citeseer, 2000.

Zaid Harchaoui, Eric Moulines, and Francis Bach. Kernel change-point analysis. *Advances in neural information processing systems*, 21, 2008.

Mikhail Hushchyn, Kenenbek Arzymatov, and Denis Derkach. Online neural networks for change-point detection. *Machine Learning*, 115, 2026.

Kaiyi Ji and Yingbin Liang. Minimax estimation of neural net distance. *Advances in Neural Information Processing Systems*, 31, 2018.

Florian Kalinke and Shakeel Gavioli-Akilagun. Optimal online change detection via random fourier features. *Advances in neural information processing systems*, 2025.

Florian Kalinke, Marco Heyden, Georg Gntuni, Edouard Fouché, and Klemens Böhm. Maximum mean discrepancy on exponential windows for online change detection. *Transactions on Machine Learning Research*, 2025.

Rohit Kanrar, Feiyu Jiang, and Zhanrui Cai. Model-free change-point detection using auc of a classifier. *Journal of Machine Learning Research*, 26(190):1–50, 2025.

Nicolas Keriven, Damien Garreau, and Iacopo Poli. NEWMA: a new method for scalable model-free online change-point detection. *IEEE Transactions on Signal Processing*, 68:3515–3528, 2020.

Vladimir Koltchinskii. Rademacher penalties and structural risk minimization. *IEEE Transactions on Information Theory*, 47(5):1902–1914, 2002.

Tze Leung Lai. Sequential changepoint detection in quality control and dynamical systems. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(4):613–644, 1995.

Yann LeCun. The MNIST database of handwritten digits. *http://yann. lecun. com/exdb/mnist/*, 1998.

Jie Li, Paul Fearnhead, Piotr Fryzlewicz, and Tengyao Wang. Automatic change-point detection in time series via deep learning. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 86(2):273–285, 2024.

Shuang Li, Yao Xie, Hanjun Dai, and Le Song. M-statistic for kernel change-point detection. *Advances in Neural Information Processing Systems*, 28, 2015.

Shuang Li, Yao Xie, Hanjun Dai, and Le Song. Scan b-statistic for kernel change-point detection. *Sequential Analysis*, 38(4):503–544, 2019.

Feng Liu, Wenkai Xu, Jie Lu, Guangquan Zhang, Arthur Gretton, and Danica J Sutherland. Learning deep kernels for non-parametric two-sample tests. In *International conference on machine learning*, pp. 6316–6326. PMLR, 2020.

Song Liu, Makoto Yamada, Nigel Collier, and Masashi Sugiyama. Change-point detection in time-series data by relative density-ratio estimation. *Neural Networks*, 43:72 – 83, 2013.

Carlos Misael Madrid Padilla, Haotian Xu, Daren Wang, Oscar Hernan Madrid Padilla, and Yi Yu. Change point detection and inference in multivariate non-parametric models under mixing conditions. *Advances in Neural Information Processing Systems*, 36, 2023.

Colin McDiarmid. On the method of bounded differences. *Surveys in combinatorics*, 141(1):148–188, 1989.

George V Moustakides and Kalliopi Basioti. Training neural networks for likelihood/density ratio estimation. *arXiv preprint arXiv:1911.00405*, 2019.

Alfred Müller. Integral probability metrics and their generating classes of functions. *Advances in applied probability*, 29(2):429–443, 1997.

Ewan S Page. Continuous inspection schemes. *Biometrika*, 41(1/2):100–115, 1954.

Victor M Panaretos and Yoav Zemel. Statistical aspects of Wasserstein distances. *Annual review of statistics and its application*, 6(1):405–431, 2019.

Nikita Puchkin and Valeriia Shcherbakova. A contrastive approach to online change point detection. In *International conference on artificial intelligence and statistics*, pp. 5686–5713. PMLR, 2023.

Mathieu Sinn, Ali Ghodsi, and Karsten Keller. Detecting change-points in time series by maximum mean discrepancy of ordinal pattern distributions. In *Proceedings of the Twenty-Eighth Conference on Uncertainty in Artificial Intelligence*, pp. 786–794, 2012.

Bharath K Sriperumbudur, Kenji Fukumizu, Arthur Gretton, Bernhard Schölkopf, and Gert RG Lanckriet. On integral probability metrics,ϕ-divergences and binary classification. *arXiv preprint arXiv:0901.2698*, 2009.

Bharath K Sriperumbudur, Kenji Fukumizu, Arthur Gretton, Bernhard Schölkopf, and Gert RG Lanckriet. Non-parametric estimation of integral probability metrics. In *2010 IEEE International Symposium on Information Theory*, pp. 1428–1432. IEEE, 2010a.

Bharath K Sriperumbudur, Arthur Gretton, Kenji Fukumizu, Bernhard Schölkopf, and Gert RG Lanckriet. Hilbert space embeddings and metrics on probability measures. *The Journal of Machine Learning Research*, 11:1517–1561, 2010b.

Alexander Stepikin, Evgenia Romanenkova, and Alexey Zaytsev. WWAggr: A window wasserstein-based aggregation for ensemble change point detection. In *2025 IEEE International Conference on Data Mining (ICDM)*, pp. 1535–1544. IEEE, 2025.

Masashi Sugiyama, Shinichi Nakajima, Hisashi Kashima, Paul Buenau, and Motoaki Kawanabe. Direct importance estimation with model selection and its application to covariate shift adaptation. *Advances in neural information processing systems*, 20, 2007.

Alexander Tartakovsky, Igor Nikiforov, and Michele Basseville. *Sequential analysis: Hypothesis testing and changepoint detection*. CRC press, 2014.

Cédric Villani et al. *Optimal transport: old and new*, volume 338. Springer, 2008.

Aladin Virmaux and Kevin Scaman. Lipschitz regularity of deep neural networks: analysis and efficient estimation. *Advances in Neural Information Processing Systems*, 31, 2018.

Xiuheng Wang, Ricardo Augusto Borsoi, Cédric Richard, and Jie Chen. Change point detection with neural online density-ratio estimator. In *IEEE international conference on acoustics, speech and signal processing (ICASSP)*, 2023.

Xiuheng Wang, Ricardo Augusto Borsoi, and Cédric Richard. Non-parametric online change point detection on riemannian manifolds. In *Forty-first International Conference on Machine Learning*, 2024.

Song Wei and Yao Xie. Online kernel cusum for change-point detection. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, pp. qkag020, 2026.

Liyan Xie, Shaofeng Zou, Yao Xie, and Venugopal V Veeravalli. Sequential (quickest) change detection: Classical results and new directions. *IEEE Journal on Selected Areas in Information Theory*, 2(2):494–514, 2021.

Liyan Xie, Yuchen Liang, and Venugopal V Veeravalli. Distributionally robust quickest change detection using Wasserstein uncertainty sets. In *International Conference on Artificial Intelligence and Statistics*, pp. 1063–1071. PMLR, 2024.

Yingyan Zeng, Yujing Huang, and Xiaoyu Chen. Beyond Euclidean summaries: Online change point detection for distribution-valued data. In *International Conference on Machine Learning*. PMLR, 2026.

Wenbin Zhou, Liyan Xie, Zhigang Peng, and Shixiang Zhu. Sequential change point detection via denoising score matching. *arXiv preprint arXiv:2501.12667*, 2025.

A Appendices

A.1 Related work

Online CPD methods can be broadly categorized into parametric and non-parametric, depending on whether prior knowledge about the data distribution is available. Parametric CPD techniques such as the cumulative sum (CUSUM) (Page, 1954; Tartakovsky et al., 2014) and the generalized likelihood ratio test (GLRT) (Gustafsson, 1996), assume that the data distribution conforms to a known parametric family. Under this assumption, they enjoy optimality guarantees on the trade-off between detection delay and false alarm rate (Lai, 1995; Tartakovsky et al., 2014), but their performance degrades when the model is misspecified. In many applications, prior knowledge of the data distribution cannot be guaranteed, leading to the development of non-parametric methods.

Early approaches monitor low-order statistics of the stream, such as its mean or variance, as in the Exponentially Weighted Moving Average (EWMA) chart (Costa & Rahim, 2006) and adaptive filtering schemes (Gustafsson, 2000). More general distributional changes are addressed by comparing the empirical distributions of pre- and post-change windows, using, for instance, nearest-neighbor graphs (Chen, 2019) or estimates of the density ratio between the two windows (Sugiyama et al., 2007; Liu et al., 2013). (Puchkin & Shcherbakova, 2023) recently cast online CPD as a sequence of contrastive classification problems.

A particularly successful line of non-parametric CPD relies on the kernel maximum mean discrepancy (MMD) (Gretton et al., 2006; 2012), an integral probability metric (Müller, 1997; Sriperumbudur et al., 2010b) that admits a closed-form empirical estimate over the unit ball of a reproducing kernel Hilbert space. The MMD was first employed for offline CPD (Harchaoui et al., 2008; Sinn et al., 2012), with subsequent work establishing model selection procedures for multiple change points (Arlot et al., 2019) and localization guarantees under mixing conditions (Madrid Padilla et al., 2023). In the online setting, Gong et al. (2012) proposed a kernelized temporal cut, while Li et al. (2015; 2019) introduced the M- and scan B-statistics, which compare a sliding post-change window against reference blocks and come with analytical thresholds derived from the asymptotic behavior of the statistic. Wei & Xie (2026) further developed an online kernel CUSUM procedure that accumulates evidence over time, and Kalinke et al. (2025) maintained the MMD over exponentially growing windows. To avoid storing historical samples, the NEWMA algorithm (Keriven et al., 2020) compares two EWMAs of random-feature embeddings computed with distinct forgetting factors, and Kalinke & Gavioli-Akilagun (2025) recently combined random Fourier features with a dyadic grid of candidate change points to obtain minimax-optimal detection with logarithmic per-step cost. Recently, non-parametric online CPD algorithms have been extended to specific non-Euclidean domains, such as graphs (Ferrari et al., 2019) and Riemannian manifolds (Wang et al., 2024).

In parallel, the Wasserstein distance from optimal transport (Villani et al., 2008) has become a central tool in statistics and machine learning (Panaretos & Zemel, 2019). Unlike kernel-based discrepancies, it takes into account the geometry of the sample space, remains informative between distributions with disjoint supports, and quantifies the magnitude of a shift rather than merely its presence. Its Kantorovich–Rubinstein dual form identifies it as the integral probability metric generated by 1-Lipschitz functions (Sriperumbudur et al., 2009; 2010a), which was exploited by Wasserstein generative adversarial networks (WGANs) (Arjovsky et al., 2017; Gulrajani et al., 2017) to approximate the distance with Lipschitz-constrained neural networks and led to considerable practical success. On the theoretical side, the concentration of the empirical Wasserstein distance is well understood (Del Barrio et al., 1999; Bolley et al., 2007; Boissard, 2011). Moreover, sample complexity results are also available for its entropic regularization (Genevay et al., 2019), and minimax rates have been derived for function classes consisting of neural networks Ji & Liang (2018). These favorable properties have motivated Wasserstein CPD. Recent work proposed to detect changes in high-dimensional streams by monitoring the Wasserstein distance between a reference batch and incoming data (Faber et al., 2021; Stepikin et al., 2025), and Ahad et al. (2025) learned the ground metric of a Sinkhorn divergence from annotated change events to perform CPD in a supervised fashion. However, the former relies on heuristic thresholds without theoretical guarantees, whereas the latter requires labeled training data, which are rarely available in practice. In a different approach, Xie et al. (2024) considered uncertainty sets in the distribution of incoming data through distributional robustness, while Zeng et al. (2026) modeled the data as discrete distributions to use an optimal transport distance.

Recently, a growing body of work seeks to combine the flexibility of neural networks with the statistical foundations of CPD. In the offline setting, (Li et al., 2024; Geng & Michailidis, 2025) trained neural networks to detect changes and quantified their error rate, and Kanrar et al. (2025) built a model-free test on the area under the ROC curve of an arbitrary classifier trained to separate the two segments, with a pivotal null distribution and localization guarantees. Cheng & Cloninger (2022) studied the influence of the geometry of the data distributions under the manifold hypothesis in classification-based two-sample testing. Online procedures have been obtained by feeding a neural classifier into a Shewhart-type chart (Hushchyn et al., 2026) or a CUSUM-type recursion (Gong et al., 2022), and by estimating the density ratio with an online neural estimator (Wang et al., 2023; Moustakides & Basioti, 2019). Zhou et al. (2025) constructed a test statistic by comparing the results of two denoising score functions evaluated on reference and test data windows.

A.2 TECHNICAL LEMMAS AND MISSING PROOFS

This section collects the proofs of the results stated in Section 4, together with the technical lemmas they rely on. For the reader's convenience, each result of Section 4 is restated (with its original numbering) before its proof.

A.2.1 PROOFS FOR SECTION 4.2

Here we prove the upper bound on $\widehat{W}_{\mathrm{NN}}[X_{1:n}|Y_{1:m}]$ under the null hypothesis (Theorem 4.5 and Corollary 4.6) and the lower bound under the contaminated alternative (Theorem 4.7 and Corollary 4.8). Lemma A.1 establishes the bounded-differences property of $\widehat{W}_{\mathrm{NN}}[X_{1:n}|Y_{1:m}]$, which allows us to apply McDiarmid's inequality in all subsequent proofs. Theorem 4.5 follows from Lemma A.2, which bounds the statistic in terms of population Rademacher complexities, and Lemma A.3, which replaces them by their empirical counterparts. Theorem 4.7 relies on Lemma A.1 together with a symmetrization argument applied to the contaminated window. Both corollaries then follow by substituting the bound of Proposition 4.4 for the Rademacher complexities.

Lemma A.1. *Under Assumptions 4.1–4.3, the empirical measures $\widehat{W}_{\mathrm{NN}}[X_{1:n}|Y_{1:m}]$ in* (10) *satisfies the bounded differences condition with*

$$c_x = \frac{2B}{n} \qquad \textit{and} \qquad c_y = \frac{2B}{m}, \tag{A.25}$$

with respect to any single coordinate $\mathbf{x}_i$ or $\mathbf{y}_j$, respectively.

Proof. Let $\widehat{W}_{\mathrm{NN}}[X_{1:n}|Y_{1:m}](\tilde{\mathbf{x}}_i)$ denote the value of (10) when $\mathbf{x}_i$ is replaced by $\tilde{\mathbf{x}}_i$ for a given index i. Then

$$\left|\widehat{W}_{\mathrm{NN}}[X_{1:n}|Y_{1:m}] - \widehat{W}_{\mathrm{NN}}[X_{1:n}|Y_{1:m}](\tilde{\mathbf{x}}_i)\right| \le \sup_{\|f\|_L \le 1} \frac{1}{n}\left|f(\mathbf{x}_i) - f(\tilde{\mathbf{x}}_i)\right| \le \frac{1}{n}\,\|\mathbf{x}_i - \tilde{\mathbf{x}}_i\| \le \frac{2B}{n}, \tag{A.26}$$

where we used Lipschitz continuity of f and $\|\mathbf{x}_i - \tilde{\mathbf{x}}_i\| \le 2B$. The same argument with a replacement of a single $\mathbf{y}_j$ yields

$$\left|\widehat{W}_{\mathrm{NN}}[X_{1:n}|Y_{1:m}] - \widehat{W}_{\mathrm{NN}}[X_{1:n}|Y_{1:m}](\tilde{\mathbf{y}}_j)\right| \le \sup_{\|f\|_L \le 1} \frac{1}{m}\left|f(\mathbf{y}_j) - f(\tilde{\mathbf{y}}_j)\right| \le \frac{1}{m}\,\|\mathbf{y}_j - \tilde{\mathbf{y}}_j\| \le \frac{2B}{m}. \tag{A.27}$$

□

Lemma A.2. *Let $X_{1:n}$ and $Y_{1:m}$ be two independent i.i.d. samples drawn from the same distribution P on $\mathcal{X}$. Under Assumptions 4.1–4.3, for any $\delta \in (0,1)$, with probability at least $1-\delta$:*

$$\widehat{W}_{\mathrm{NN}}[X_{1:n}|Y_{1:m}] \;\le\; 2\Big(\mathcal{R}_{P,n}(\mathcal{F}_{\mathrm{NN}}) + \mathcal{R}_{P,m}(\mathcal{F}_{\mathrm{NN}})\Big) + \sqrt{\frac{c_{xy}^2\,\ln(1/\delta)}{2}}, \tag{A.28}$$

where $c_x = \frac{2B}{n}$, $c_y = \frac{2B}{m}$, $c_{xy}^2 := nc_x^2 + mc_y^2 = 4B^2\big(\frac{1}{n} + \frac{1}{m}\big)$.

Proof. By Lemma A.1, $\widehat{W}_{\mathrm{NN}}[X_{1:n}|Y_{1:m}]$ satisfies the bounded-differences condition with $c_x = \frac{2B}{n}$ and $c_y = \frac{2B}{m}$. McDiarmid's inequality then gives, with probability at least $1-\delta$,

$$\widehat{W}_{\mathrm{NN}}[X_{1:n}|Y_{1:m}] \le \mathbb{E}_P\big[\widehat{W}_{\mathrm{NN}}[X_{1:n}|Y_{1:m}]\big] + \sqrt{\frac{c_{xy}^2\,\ln(1/\delta)}{2}}. \tag{A.29}$$

It remains to bound the expectation $\mathbb{E}_P\big[\widehat{W}_{\mathrm{NN}}[X_{1:n}|Y_{1:m}]\big]$ in terms of the Rademacher complexities of $\mathcal{F}_{\mathrm{NN}}$. By the definition in (10), we have

$$\begin{aligned}
&\mathbb{E}_P\Big[\widehat{W}_{\mathrm{NN}}[X_{1:n}|Y_{1:m}]\Big] \\
&= \mathbb{E}_P\left[\sup_{f\in\mathcal{F}_{\mathrm{NN}}}\left\{\frac{1}{n}\sum_{i=1}^{n} f(\mathbf{x}_i) - \frac{1}{m}\sum_{j=1}^{m} f(\mathbf{y}_j)\right\}\right] \\
&\overset{(a)}{=} \mathbb{E}_P\left[\sup_{f\in\mathcal{F}_{\mathrm{NN}}}\left\{\frac{1}{n}\sum_{i=1}^{n} f(\mathbf{x}_i) - \mathbb{E}_P\{f(\mathbf{x}'_i)\} - \frac{1}{m}\sum_{j=1}^{m} f(\mathbf{y}_j) + \mathbb{E}_P\{f(\mathbf{y}'_j)\}\right\}\right] \\
&\overset{(b)}{\leq} \mathbb{E}_P\left[\sup_{f\in\mathcal{F}_{\mathrm{NN}}}\left\{\frac{1}{n}\sum_{i=1}^{n}\big[f(\mathbf{x}_i) - f(\mathbf{x}'_i)\big] - \frac{1}{m}\sum_{j=1}^{m}\big[f(\mathbf{y}_j) - f(\mathbf{y}'_j)\big]\right\}\right] \\
&\overset{(c)}{=} \mathbb{E}_{\sigma,P}\left[\sup_{f\in\mathcal{F}_{\mathrm{NN}}}\left\{\frac{1}{n}\sum_{i=1}^{n}\sigma_i\big[f(\mathbf{x}_i) - f(\mathbf{x}'_i)\big] - \frac{1}{m}\sum_{j=1}^{m}\sigma_j\big[f(\mathbf{y}'_j) - f(\mathbf{y}_j)\big]\right\}\right] \\
&\overset{(d)}{\leq} \mathbb{E}_{\sigma,P}\left[\sup_{f\in\mathcal{F}_{\mathrm{NN}}}\left\{\frac{1}{n}\sum_{i=1}^{n}\sigma_i\big[f(\mathbf{x}_i) - f(\mathbf{x}'_i)\big]\right\} + \sup_{f\in\mathcal{F}_{\mathrm{NN}}}\left\{\frac{1}{m}\sum_{j=1}^{m}\sigma_j\big[f(\mathbf{y}'_j) - f(\mathbf{y}_j)\big]\right\}\right] \\
&\overset{(e)}{\leq} 2\,\mathbb{E}_{\sigma,P}\left[\sup_{f\in\mathcal{F}_{\mathrm{NN}}}\left\{\frac{1}{n}\sum_{i=1}^{n}\sigma_i f(\mathbf{x}_i)\right\} + \sup_{f\in\mathcal{F}_{\mathrm{NN}}}\left\{\frac{1}{m}\sum_{j=1}^{m}\sigma_j f(\mathbf{y}_j)\right\}\right] \\
&= 2\Big(\mathcal{R}_{P,n}(\mathcal{F}_{\mathrm{NN}}) + \mathcal{R}_{P,m}(\mathcal{F}_{\mathrm{NN}})\Big).
\end{aligned} \tag{A.30}$$

Here, step (a) introduces independent ghost samples $\mathbf{x}'_i$ and $\mathbf{y}'_j$ (both also drawn from P) and centers each empirical sum by its population mean under $P = Q$. Step (b) exchanges the corresponding expectations with the supremum. This enables the Rademacher symmetrization carried out in (c), where $\sigma_1, \ldots, \sigma_n$ denote i.i.d. Rademacher random variables uniformly distributed on $\{-1, 1\}$. Step (d) uses subadditivity of the supremum to separate the two terms. Step (e) uses the exchangeability of the ghost-sample pairs $(\mathbf{x}_i, \mathbf{x}'_i)$ and $(\mathbf{y}_j, \mathbf{y}'_j)$ to relate each symmetrized difference to twice the standard Rademacher process. □

Lemma A.3. *Let $X_{1:n}$ be i.i.d. samples drawn from a distribution P on $\mathcal{X}$. Under Assumptions 4.1–4.3, for any $\delta \in (0, 1)$, with probability at least $1 - \delta$,*

$$\mathcal{R}_{P,n}(\mathcal{F}_{\mathrm{NN}}) \leq \hat{\mathcal{R}}_n(\mathcal{F}_{\mathrm{NN}}) + B\sqrt{\frac{2\ln(1/\delta)}{n}}. \tag{A.31}$$

Proof. We show that $\hat{\mathcal{R}}_n(\mathcal{F}_{\mathrm{NN}})$ satisfies bounded differences with respect to replacing a single sample $\mathbf{x}_i$ by $\tilde{\mathbf{x}}_i$. Let $\hat{\mathcal{R}}_n(\mathcal{F}_{\mathrm{NN}})(\tilde{\mathbf{x}}_i)$ denote the same quantity with $\mathbf{x}_i$ replaced by $\tilde{\mathbf{x}}_i$. Then

$$\begin{aligned}
\big|\hat{\mathcal{R}}_n(\mathcal{F}_{\mathrm{NN}}) - \hat{\mathcal{R}}_n(\mathcal{F}_{\mathrm{NN}})(\tilde{\mathbf{x}}_i)\big| &\overset{(a)}{\leq} \mathbb{E}_\sigma\left[\sup_{f\in\mathcal{F}_{\mathrm{NN}}}\frac{1}{n}\big|\sigma_i\big(f(\mathbf{x}_i) - f(\tilde{\mathbf{x}}_i)\big)\big|\right] \\
&\overset{(b)}{\leq} \frac{1}{n}\sup_{f\in\mathcal{F}_{\mathrm{NN}}}\big|f(\mathbf{x}_i) - f(\tilde{\mathbf{x}}_i)\big| \\
&\overset{(c)}{\leq} \frac{1}{n}\sup_{f\in\mathcal{F}_{\mathrm{NN}}}\|f\|_L\,\|\mathbf{x}_i - \tilde{\mathbf{x}}_i\| \\
&\overset{(d)}{\leq} \frac{2B}{n},
\end{aligned} \tag{A.32}$$

where step (a) follows from the fact that the samples $X_{1:n}$ and $\mathbf{x}'_{1:n}$ differ only at index i, step (b) relies on the fact that $|\sigma_i| = 1$, step (c) uses the Lipschitz continuity of f, and step (d) uses $\|f\|_L \leq 1$

from Assumption 4.2 and $\|\mathbf{x}_i - \tilde{\mathbf{x}}_i\| \le \|\mathbf{x}_i\| + \|\tilde{\mathbf{x}}_i\| \le 2B$. Thus, the bounded-differences constants are $c_i = \frac{2B}{n}$ for all i. By McDiarmid's inequality, with probability at least $1-\delta$,

$$\mathcal{R}_{P,n}(\mathcal{F}_{\mathrm{NN}}) - \hat{\mathcal{R}}_n(\mathcal{F}_{\mathrm{NN}}) \le \sqrt{\frac{1}{2}\sum_{i=1}^{n} c_i^2 \,\ln(1/\delta)} = B\sqrt{\frac{2\ln(1/\delta)}{n}}. \tag{A.33}$$

□

Theorem 4.5. *Let $X_{1:n}$ and $Y_{1:m}$ be two independent samples, each consisting of i.i.d. entries drawn from the same distribution P on $\mathcal{X}$. Under Assumptions 4.1–4.3, for any $\delta \in (0,1)$, with probability at least $1-\delta$,*

$$\widehat{W}_{\mathrm{NN}}[X_{1:n}|Y_{1:m}] \;\le\; 2\Big(\hat{\mathcal{R}}_n(\mathcal{F}_{\mathrm{NN}}) + \hat{\mathcal{R}}_m(\mathcal{F}_{\mathrm{NN}})\Big) + 3\sqrt{2}\,B\Big(\frac{1}{\sqrt{n}} + \frac{1}{\sqrt{m}}\Big)\sqrt{\ln\Big(\frac{3}{\delta}\Big)}. \tag{19}$$

Proof. By Lemma A.2, for any $\delta_1 \in (0,1)$, with probability at least $1-\delta_1$,

$$\widehat{W}_{\mathrm{NN}}[X_{1:n}|Y_{1:m}] \le 2\Big(\mathcal{R}_{P,n}(\mathcal{F}_{\mathrm{NN}}) + \mathcal{R}_{P,m}(\mathcal{F}_{\mathrm{NN}})\Big) + \sqrt{\frac{c_{xy}^2\ln(1/\delta_1)}{2}}, \tag{A.34}$$

where $c_{xy}^2 = 4B^2\big(\frac{1}{n}+\frac{1}{m}\big)$ from Lemma A.1. Applying Lemma A.3 to $X_{1:n}$ and $Y_{1:m}$, for any $\delta_2, \delta_3 \in (0,1)$, yields

$$\mathcal{R}_{P,n}(\mathcal{F}_{\mathrm{NN}}) \le \hat{\mathcal{R}}_n(\mathcal{F}_{\mathrm{NN}}) + B\sqrt{\frac{2\ln(1/\delta_2)}{n}}, \qquad \mathcal{R}_{P,m}(\mathcal{F}_{\mathrm{NN}}) \le \hat{\mathcal{R}}_m(\mathcal{F}_{\mathrm{NN}}) + B\sqrt{\frac{2\ln(1/\delta_3)}{m}}. \tag{A.35}$$

Choosing $\delta_1 = \delta_2 = \delta_3 = \delta/3$ and using the union bound, with probability at least $1-\delta$,

$$\begin{aligned}\widehat{W}_{\mathrm{NN}}[X_{1:n}|Y_{1:m}] &\le 2\Big(\hat{\mathcal{R}}_n(\mathcal{F}_{\mathrm{NN}}) + \hat{\mathcal{R}}_m(\mathcal{F}_{\mathrm{NN}})\Big) + 2B\sqrt{2\ln\Big(\frac{3}{\delta}\Big)}\Big(\frac{1}{\sqrt{n}}+\frac{1}{\sqrt{m}}\Big) \\ &\quad + \sqrt{2B^2\left(\frac{1}{n}+\frac{1}{m}\right)\ln\Big(\frac{3}{\delta}\Big)} \\ &\le 2\Big(\hat{\mathcal{R}}_n(\mathcal{F}_{\mathrm{NN}}) + \hat{\mathcal{R}}_m(\mathcal{F}_{\mathrm{NN}})\Big) + 3\sqrt{2}\,B\Big(\frac{1}{\sqrt{n}}+\frac{1}{\sqrt{m}}\Big)\sqrt{\ln\Big(\frac{3}{\delta}\Big)},\end{aligned} \tag{A.36}$$

where we used $\sqrt{\frac{1}{n}+\frac{1}{m}} \le \frac{1}{\sqrt{n}} + \frac{1}{\sqrt{m}}$ to combine the last two terms. □

Corollary 4.6. *Let $X_{1:n}$ and $Y_{1:m}$ be two independent samples, each consisting of i.i.d. entries drawn from the same distribution P on $\mathcal{X}$. Under Assumptions 4.1–4.3, for any $\delta \in (0,1)$, with probability at least $1-\delta$,*

$$\widehat{W}_{\mathrm{NN}}[X_{1:n}|Y_{1:m}] \;\le\; B\Big(\frac{1}{\sqrt{n}}+\frac{1}{\sqrt{m}}\Big)\left[2\prod_{\ell=1}^{d} M_\ell\big(\sqrt{2\log(2)\,d}+1\big) + 3\sqrt{2\,\ln\Big(\frac{3}{\delta}\Big)}\right]. \tag{20}$$

Proof. By Theorem 4.5,

$$\widehat{W}_{\mathrm{NN}}[X_{1:n}|Y_{1:m}] \le 2\Big(\hat{\mathcal{R}}_n(\mathcal{F}_{\mathrm{NN}}) + \hat{\mathcal{R}}_m(\mathcal{F}_{\mathrm{NN}})\Big) + 3\sqrt{2}\,B\Big(\frac{1}{\sqrt{n}}+\frac{1}{\sqrt{m}}\Big)\sqrt{\ln\Big(\frac{3}{\delta}\Big)}. \tag{A.37}$$

Applying Proposition 4.4 to both $\hat{\mathcal{R}}_n(\mathcal{F}_{\mathrm{NN}})$ and $\hat{\mathcal{R}}_m(\mathcal{F}_{\mathrm{NN}})$ yields

$$2\Big(\hat{\mathcal{R}}_n(\mathcal{F}_{\mathrm{NN}}) + \hat{\mathcal{R}}_m(\mathcal{F}_{\mathrm{NN}})\Big) \le 2B\prod_{\ell=1}^{d} M_\ell\Big(\frac{1}{\sqrt{n}}+\frac{1}{\sqrt{m}}\Big)\big(\sqrt{2\log(2)\,d}+1\big), \tag{A.38}$$

which proves the claim. □

Theorem 4.7. *Let P and Q be two probability distributions supported on $\mathcal{X}$, and let $X_{1:n} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ and $Y_{1:m} = \{\mathbf{y}_1, \ldots, \mathbf{y}_m\}$ be two sets of independent samples distributed as follows:*

- *the samples in $X_{1:n}$ are drawn i.i.d. from distribution P.*
- *the samples in $Y_{1:m}$ are drawn as: $\mathbf{y}_1, \ldots, \mathbf{y}_{(1-\epsilon)m}$ be drawn i.i.d. from distribution P, while $\mathbf{y}_{(1-\epsilon)m+1}, \ldots, \mathbf{y}_m$ be drawn i.i.d. from distribution Q.*

where the "degree of contamination" $\epsilon \in (0, 1]$ such that ϵm is an integer. Under Assumptions 4.1–4.3, for any $\delta \in (0, 1)$, with probability at least $1 - \delta$:

$$\widehat{W}_{\mathrm{NN}}[X_{1:n}|Y_{1:m}] \geq \epsilon\, W_{\mathrm{NN}}(P, Q) - 2\mathcal{R}_{P,n}(\mathcal{F}_{\mathrm{NN}}) - 2\mathcal{R}_{Y,m}(\mathcal{F}_{\mathrm{NN}}) - \sqrt{\frac{c_{xy}^2 \ln(1/\delta)}{2}}, \tag{21}$$

where $c_{xy}^2 := 4B^2\big(\frac{1}{n} + \frac{1}{m}\big)$, and $\mathcal{R}_{Y,m}(\mathcal{F}_{\mathrm{NN}})$ denotes the Rademacher complexity of the contaminated second window, i.e., computed with the expectation taken over the product law $P^{\otimes(1-\epsilon)m} \otimes Q^{\otimes \epsilon m}$ of the samples $Y_{1:m}$.

Proof. By Lemma A.1, $\widehat{W}_{\mathrm{NN}}[X_{1:n}|Y_{1:m}]$ satisfies the bounded-differences condition with $c_x = \frac{2B}{n}$ and $c_y = \frac{2B}{m}$. Thus, McDiarmid's inequality implies that with probability at least $1 - \delta$,

$$\widehat{W}_{\mathrm{NN}}[X_{1:n}|Y_{1:m}] \geq \mathbb{E}\big[\widehat{W}_{\mathrm{NN}}[X_{1:n}|Y_{1:m}]\big] - \sqrt{\frac{c_{xy}^2 \ln(1/\delta)}{2}}. \tag{A.39}$$

It remains to lower bound the expectation $\mathbb{E}\big[\widehat{W}_{\mathrm{NN}}[X_{1:n}|Y_{1:m}]\big]$ (note that the expectation is taken w.r.t. the joint distribution of the data $X_{1:n}$ and $Y_{1:m}$). By the definition of the estimated Wasserstein-1 distance via Kantorovich–Rubinstein duality in (10), this term can be written as:

$$\mathbb{E}\left[\widehat{W}_{\mathrm{NN}}[X_{1:n}|Y_{1:m}]\right] = \mathbb{E}\left[\sup_{f \in \mathcal{F}_{\mathrm{NN}}} \left\{ \frac{1}{n}\sum_{i=1}^{n} f(\mathbf{x}_i) - \frac{1}{m}\sum_{j=1}^{m} f(\mathbf{y}_j) \right\}\right]. \tag{A.40}$$

We decompose the expression inside the supremum inside the expectation in the above equation as follows:

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n} f(\mathbf{x}_i) - \frac{1}{m}\sum_{j=1}^{m} f(\mathbf{y}_j) = &\left(\frac{1}{n}\sum_{i=1}^{n} f(\mathbf{x}_i) - \mathbb{E}_X[f(\mathbf{x})]\right) \\ &- \left(\frac{1}{m}\sum_{j=1}^{m} f(\mathbf{y}_j) - \mathbb{E}_Y[f(\mathbf{y})]\right) \\ &+ \mathbb{E}_X[f(\mathbf{x})] - \mathbb{E}_Y[f(\mathbf{y})],\end{aligned}$$

where $\mathbf{x} \sim P$ denotes a random variable distributed according to P, i.e.,

$$\mathbb{E}_X[f(\mathbf{x})] = \mathbb{E}_P[f(\mathbf{x})]\,, \tag{A.41}$$

while $\mathbf{y}$ denotes a random variable distributed according to the mixture distribution $\mathbf{y} \sim (1-\epsilon)P + \epsilon Q$, which implies that

$$\mathbb{E}_Y[f(\mathbf{y})] = (1-\epsilon)\mathbb{E}_P[f(\mathbf{y})] + \epsilon\mathbb{E}_Q[f(\mathbf{y})]\,. \tag{A.42}$$

Note that the mixture expectation coincides exactly with the average of the per-sample expectations of the second window,

$$\mathbb{E}_Y[f(\mathbf{y})] = \frac{1}{m}\sum_{j=1}^{m} \mathbb{E}[f(\mathbf{y}_j)]\,, \tag{A.43}$$

since $(1-\epsilon)m$ of the samples have mean $\mathbb{E}_P[f]$ and ϵm of them have mean $\mathbb{E}_Q[f]$. This identity is exact and does not require the samples $\mathbf{y}_j$ to be identically distributed: the statistic only uses the empirical average of the second window, so the ordering of the samples plays no role.

For convenience, let us denote:

$$A(f) = \frac{1}{n}\sum_{i=1}^{n} f(\mathbf{x}_i) - \mathbb{E}_P[f(\mathbf{x})], \tag{A.44}$$

$$B(f) = \frac{1}{m}\sum_{j=1}^{m} f(\mathbf{y}_j) - \mathbb{E}_Y[f(\mathbf{y})], \tag{A.45}$$

$$C(f) = \mathbb{E}_P[f(\mathbf{x})] - \mathbb{E}_Y[f(\mathbf{y})]. \tag{A.46}$$

Thus, the expression we intend to bound can be written as:

$$\mathbb{E}_{X,Y}\left[\sup_{f\in\mathcal{F}_{\mathrm{NN}}}\{A(f)-B(f)+C(f)\}\right]. \tag{A.47}$$

Using properties of the supremum, we can lower bound the term inside the expectation as:

$$\sup_{f\in\mathcal{F}_{\mathrm{NN}}}\{A(f)-B(f)+C(f)\}\geq \sup_{f\in\mathcal{F}_{\mathrm{NN}}}\{C(f)\}+\inf_{f\in\mathcal{F}_{\mathrm{NN}}}\{A(f)-B(f)\} \tag{A.48}$$

$$\geq \sup_{f\in\mathcal{F}_{\mathrm{NN}}}\{C(f)\}-\sup_{f\in\mathcal{F}_{\mathrm{NN}}}\{-A(f)\}-\sup_{f\in\mathcal{F}_{\mathrm{NN}}}\{B(f)\} \tag{A.49}$$

where the first inequality holds since $A(f)-B(f)\geq \inf_{f'\in\mathcal{F}_{\mathrm{NN}}}\{A(f')-B(f')\}$ for every f, so that $C(f)+\inf_{f'\in\mathcal{F}_{\mathrm{NN}}}\{A(f')-B(f')\}$ is a minorizer of $A(f)-B(f)+C(f)$; and the second follows from $\inf_f\{A(f)-B(f)\}=-\sup_f\{B(f)-A(f)\}\geq -\sup_f\{-A(f)\}-\sup_f\{B(f)\}$, using subadditivity of the supremum. Taking expectations on both sides, we get:

$$\mathbb{E}_{X,Y}\left[\sup_{f\in\mathcal{F}_{\mathrm{NN}}}\{A(f)-B(f)+C(f)\}\right]\geq \sup_{f\in\mathcal{F}_{\mathrm{NN}}}\{C(f)\}-\mathbb{E}_P\left[\sup_{f\in\mathcal{F}_{\mathrm{NN}}}\{-A(f)\}\right]$$

$$-\mathbb{E}_Y\left[\sup_{f\in\mathcal{F}_{\mathrm{NN}}}\{B(f)\}\right]. \tag{A.50}$$

We now bound each of these terms.

First term: The term $\sup_{f\in\mathcal{F}_{\mathrm{NN}}}\{C(f)\}$ can be computed exactly using the fact that $\mathbf{y}$ comes from the mixture distribution $(1-\epsilon)P+\epsilon Q$:

$$\begin{aligned}\sup_{f\in\mathcal{F}_{\mathrm{NN}}}\{C(f)\}&=\sup_{f\in\mathcal{F}_{\mathrm{NN}}}\Big\{\mathbb{E}_P[f(\mathbf{x})]-\mathbb{E}_Y[f(\mathbf{y})]\Big\}\\&=\sup_{f\in\mathcal{F}_{\mathrm{NN}}}\Big\{\mathbb{E}_P[f(\mathbf{x})]-(1-\epsilon)\,\mathbb{E}_P[f(\mathbf{y})]-\epsilon\,\mathbb{E}_Q[f(\mathbf{y})]\Big\}\\&=\epsilon\left(\sup_{f\in\mathcal{F}_{\mathrm{NN}}}\mathbb{E}_P[f(\mathbf{x})]-\mathbb{E}_Q[f(\mathbf{y})]\right)\\&=\epsilon\,W_{\mathrm{NN}}(P,Q),\end{aligned} \tag{A.51}$$

where the last equality is the definition of the neural-network Wasserstein distance in (9). This shows the influence of the level of contamination in the window and of the neural-network Wasserstein distance between the pre- and post-change distributions.

Second term: We upper bound $\mathbb{E}_P\left[\sup_{f\in\mathcal{F}_{\mathrm{NN}}}\{-A(f)\}\right]$. Since $f\in\mathcal{F}_{\mathrm{NN}}$ implies $-f\in\mathcal{F}_{\mathrm{NN}}$, substituting $f\mapsto -f$ yields

$$\sup_{f\in\mathcal{F}_{\mathrm{NN}}}\{-A(f)\}=\sup_{f\in\mathcal{F}_{\mathrm{NN}}}\left\{\mathbb{E}_P[f(\mathbf{x})]-\frac{1}{n}\sum_{i=1}^n f(\mathbf{x}_i)\right\}=\sup_{f\in\mathcal{F}_{\mathrm{NN}}}\{A(f)\}. \tag{A.52}$$

The expectation of the latter can be bounded by ghost-sample symmetrization:

$$\begin{aligned}\mathbb{E}_P\left[\sup_{f\in\mathcal{F}_{\mathrm{NN}}}\{A(f)\}\right]&\overset{(a)}{=}\mathbb{E}_P\left[\sup_{f\in\mathcal{F}_{\mathrm{NN}}}\frac{1}{n}\sum_{i=1}^n\Big(f(\mathbf{x}_i)-\mathbb{E}_P\{f(\mathbf{x}'_i)\}\Big)\right]\\&\overset{(b)}{\leq}\mathbb{E}_P\left[\sup_{f\in\mathcal{F}_{\mathrm{NN}}}\frac{1}{n}\sum_{i=1}^n\Big(f(\mathbf{x}_i)-f(\mathbf{x}'_i)\Big)\right]\\&\overset{(c)}{=}\mathbb{E}_{P,\boldsymbol{\sigma}}\left[\sup_{f\in\mathcal{F}_{\mathrm{NN}}}\frac{1}{n}\sum_{i=1}^n\sigma_i\Big(f(\mathbf{x}_i)-f(\mathbf{x}'_i)\Big)\right]\\&\overset{(d)}{\leq}\mathbb{E}_{P,\boldsymbol{\sigma}}\left[\sup_{f\in\mathcal{F}_{\mathrm{NN}}}\frac{1}{n}\sum_{i=1}^n\sigma_i f(\mathbf{x}_i)\right]+\mathbb{E}_{P,\boldsymbol{\sigma}}\left[\sup_{f\in\mathcal{F}_{\mathrm{NN}}}\frac{1}{n}\sum_{i=1}^n\sigma_i f(\mathbf{x}'_i)\right]\\&\overset{(e)}{=}2\,\mathcal{R}_{P,n}(\mathcal{F}_{\mathrm{NN}}),\end{aligned} \tag{A.53}$$

where step (a) introduces i.i.d. ghost samples $\mathbf{x}'_1, \dots, \mathbf{x}'_n \sim P$ independent of $X_{1:n}$, so that $\mathbb{E}_P\{f(\mathbf{x}'_i)\} = \mathbb{E}_P\{f(\mathbf{x})\}$; step (b) exchanges the expectation over the ghost samples with the supremum (Jensen's inequality); step (c) uses the exchangeability of the pairs $(\mathbf{x}_i, \mathbf{x}'_i)$, under which the joint distribution of $\{f(\mathbf{x}_i) - f(\mathbf{x}'_i)\}_{i=1}^n$ is unchanged when multiplied by the Rademacher variables σ_i; step (d) uses subadditivity of the supremum together with the fact that $-\sigma_i$ has the same distribution as σ_i; and step (e) follows from the definition of the Rademacher complexity, since $\mathbf{x}_i$ and $\mathbf{x}'_i$ are both i.i.d. samples from P. We conclude that

$$\mathbb{E}_P\left[\sup_{f\in\mathcal{F}_{\mathrm{NN}}}\{-A(f)\}\right] \le 2\,\mathcal{R}_{P,n}(\mathcal{F}_{\mathrm{NN}})\,. \tag{A.54}$$

Third term: For this term we need to upper bound $\mathbb{E}_Y\left[\sup_{f\in\mathcal{F}_{\mathrm{NN}}}\{B(f)\}\right]$. The samples $\mathbf{y}_1, \dots, \mathbf{y}_m$ are independent but not identically distributed: $\mathbf{y}_j \sim P_j$ with $P_j = P$ for $j \le (1-\epsilon)m$ and $P_j = Q$ for $j > (1-\epsilon)m$. Using the identity $\mathbb{E}_Y[f(\mathbf{y})] = \frac{1}{m}\sum_{j=1}^m \mathbb{E}[f(\mathbf{y}_j)]$ established above, $B(f)$ can be written exactly as a normalized sum of centered independent terms,

$$B(f) = \frac{1}{m}\sum_{j=1}^m \Big(f(\mathbf{y}_j) - \mathbb{E}[f(\mathbf{y}_j)]\Big)\,. \tag{A.55}$$

We can therefore apply the same ghost-sample symmetrization as for the second term, which remains valid for independent, non-identically distributed samples: introducing ghost samples $\mathbf{y}'_j \sim P_j$, independent of each other and of $Y_{1:m}$,

$$\begin{aligned}
\mathbb{E}_Y\left[\sup_{f\in\mathcal{F}_{\mathrm{NN}}}\{B(f)\}\right] &\overset{(a)}{=} \mathbb{E}_Y\left[\sup_{f\in\mathcal{F}_{\mathrm{NN}}}\frac{1}{m}\sum_{j=1}^m\Big(f(\mathbf{y}_j) - \mathbb{E}\{f(\mathbf{y}'_j)\}\Big)\right]\\
&\overset{(b)}{\le} \mathbb{E}_Y\left[\sup_{f\in\mathcal{F}_{\mathrm{NN}}}\frac{1}{m}\sum_{j=1}^m\Big(f(\mathbf{y}_j) - f(\mathbf{y}'_j)\Big)\right]\\
&\overset{(c)}{=} \mathbb{E}_{Y,\boldsymbol{\sigma}}\left[\sup_{f\in\mathcal{F}_{\mathrm{NN}}}\frac{1}{m}\sum_{j=1}^m\sigma_j\Big(f(\mathbf{y}_j) - f(\mathbf{y}'_j)\Big)\right]\\
&\overset{(d)}{\le} \mathbb{E}_{Y,\boldsymbol{\sigma}}\left[\sup_{f\in\mathcal{F}_{\mathrm{NN}}}\frac{1}{m}\sum_{j=1}^m\sigma_j f(\mathbf{y}_j)\right] + \mathbb{E}_{Y,\boldsymbol{\sigma}}\left[\sup_{f\in\mathcal{F}_{\mathrm{NN}}}\frac{1}{m}\sum_{j=1}^m\sigma_j f(\mathbf{y}'_j)\right]\\
&\overset{(e)}{=} 2\,\mathcal{R}_{Y,m}(\mathcal{F}_{\mathrm{NN}})\,,
\end{aligned} \tag{A.56}$$

where step (a) uses $\mathbb{E}\{f(\mathbf{y}'_j)\} = \mathbb{E}\{f(\mathbf{y}_j)\}$ (each ghost has the same law P_j as its counterpart); step (b) exchanges the expectation over the ghost samples with the supremum (Jensen's inequality); step (c) uses the exchangeability of each pair $(\mathbf{y}_j, \mathbf{y}'_j)$ together with the independence of the pairs across j; step (d) uses subadditivity of the supremum together with $-\sigma_j$ having the same distribution as σ_j; and step (e) is the definition of the Rademacher complexity of the contaminated sample, with the expectation taken over the product law $P^{\otimes(1-\epsilon)m} \otimes Q^{\otimes\epsilon m}$, as in the statement of the theorem.

Combining the bounds on the three terms with the McDiarmid step at the beginning of the proof completes the argument. □

Corollary 4.8. *Let P, Q, $X_{1:n}$, and $Y_{1:m}$ be as in Theorem 4.7. Under Assumptions 4.1–4.3, for any $\delta \in (0,1)$, with probability at least $1-\delta$,*

$$\begin{aligned}
\widehat{W}_{\mathrm{NN}}[X_{1:n}\,|\,Y_{1:m}] \;\ge\; &\epsilon\, W_{\mathrm{NN}}(P,Q) - 2B\prod_{\ell=1}^d M_\ell\big(\sqrt{2\log(2)\,d}+1\big)\left(\frac{1}{\sqrt{n}}+\frac{1}{\sqrt{m}}\right)\\
&- \sqrt{2}\,B\sqrt{\Big(\frac{1}{n}+\frac{1}{m}\Big)\ln\Big(\frac{1}{\delta}\Big)}\,.
\end{aligned} \tag{22}$$

Proof. Since the upper bound in Proposition 4.4 holds uniformly over the data points in $\mathcal{X}$, it also bounds the population Rademacher complexities regardless of the underlying distributions, namely $\mathcal{R}_{P,n}(\mathcal{F}_{\mathrm{NN}}) \le B\prod_{\ell=1}^{d} M_\ell(\sqrt{2\log(2)\,d}+1)/\sqrt{n}$ and $\mathcal{R}_{Y,m}(\mathcal{F}_{\mathrm{NN}}) \le B\prod_{\ell=1}^{d} M_\ell(\sqrt{2\log(2)\,d}+1)/\sqrt{m}$. Substituting these bounds into Theorem 4.7 and using $c_{xy}^2 = 4B^2\left(\frac{1}{n}+\frac{1}{m}\right)$ to rewrite $\sqrt{\frac{c_{xy}^2\ln(1/\delta)}{2}} = \sqrt{2}\,B\sqrt{\left(\frac{1}{n}+\frac{1}{m}\right)\ln(1/\delta)}$ proves the claim. □

A.2.2 Proofs for Section 4.3

Here we prove the ARL bound (Theorem 4.9), the FAP bound (Theorem 4.10), and the detection delay bound (Theorem 4.11). Theorems 4.9 and 4.10 both rely on the per-test tail bound provided by Corollary 4.6, combined with a union bound over the alarm events that does not require independence across tests; Theorem 4.9 additionally uses the tail-sum formula for the expectation of the stopping time. Theorem 4.11 follows from Lemma A.4, which bounds the probability of a missed detection at a single window position using Corollary 4.8, together with the observation that the event $\{D \ge k\}$ is contained in the miss event at the k-th post-change window position.

Theorem 4.9 (Average run length). *Let T be the stopping time* (23) *with a constant threshold $\nu_t \equiv \nu$, and let $\gamma \ge \frac{1}{2}$. Under Assumptions 4.1–4.3, if ν is chosen such that $\nu \ge C_{n,m}\big(C_d + 3\sqrt{2}\,\sqrt{\ln(6\gamma)}\big)$, where $C_{n,m} := B\big(\frac{1}{\sqrt{n}}+\frac{1}{\sqrt{m}}\big)$ and $C_d := 2\prod_{\ell=1}^{d} M_\ell\big(\sqrt{2\log(2)\,d}+1\big)$, then $\mathbb{E}_\infty[T] \ge \gamma$.*

Proof. The argument combines an exponential tail bound on the alarm probability at each test, obtained from Corollary 4.6, with a union bound over the alarm events and the tail-sum formula for $\mathbb{E}_\infty[T]$. For every fixed t, the two windows $X^{(t)}$ and $Y^{(t)}$ in (11) are disjoint, hence independent, and each consists of i.i.d. samples from P; therefore Corollary 4.6 applies to $g_t = \widehat{W}_{\mathrm{NN}}[X^{(t)} \mid Y^{(t)}]$ and yields, for any $\delta \in (0,1)$,

$$g_t \le C_{n,m}\left(C_d + 3\sqrt{2}\,\sqrt{\ln(3/\delta)}\right) \qquad \text{with probability at least } 1-\delta. \tag{A.57}$$

Setting $\delta = 3\exp(-((\nu_t - C_{n,m}C_d)/(3\sqrt{2}C_{n,m}))^2)$, which lies in $(0,1]$ whenever $\nu_t \ge C_{n,m}C_d$, gives the tail bound

$$\mathbb{P}_\infty\left(g_t > \nu_t\right) \le 3\exp\left(-\left(\frac{\nu_t - C_{n,m}C_d}{3\sqrt{2}\,C_{n,m}}\right)^2\right), \qquad \nu_t \ge C_{n,m}C_d. \tag{A.58}$$

Let $A_t := \{g_t > \nu_t\}$ denote the event of an alarm at time t. If the threshold is constant over time ($\nu_t \equiv \nu$ with $\nu \ge C_{n,m}C_d$), then (A.58) yields the per-test false-alarm bound

$$\mathbb{P}(\nu) := \mathbb{P}_\infty(A_t) \le 3\exp\left(-\left(\frac{\nu - C_{n,m}C_d}{3\sqrt{2}\,C_{n,m}}\right)^2\right), \tag{A.59}$$

where $\mathbb{P}_\infty(A_t)$ does not depend on t because the stream is i.i.d. under the global null, so that g_t has the same distribution for every t. Without assuming independence across tests, the union bound gives, for every integer $k \ge 1$,

$$\mathbb{P}_\infty(T \le k) = \mathbb{P}_\infty\left(\bigcup_{t=1}^{k} A_t\right) \le \sum_{t=1}^{k}\mathbb{P}_\infty(A_t) = k\,\mathbb{P}(\nu). \tag{A.60}$$

Hence $\mathbb{P}_\infty(T > k) \ge 1 - k\,\mathbb{P}(\nu)$ for all k such that $k\,\mathbb{P}(\nu) \le 1$, and trivially $\mathbb{P}_\infty(T > k) \ge 0$ otherwise. Using the tail-sum formula for the integer-valued stopping time $T \ge 1$ in (23),

$$\mathbb{E}_\infty[T] = \sum_{k=0}^{\infty}\mathbb{P}_\infty(T > k) \ge \sum_{k=0}^{\infty}(1 - k\,\mathbb{P}(\nu))_+, \tag{A.61}$$

and since $h(x) := (1 - x\,\mathbb{P}(\nu))_+$ is nonnegative and decreasing on $\mathbb{R}_+$,

$$\mathbb{E}_\infty[T] \ge \int_0^{1/\mathbb{P}(\nu)} h(x)\,dx = \frac{1}{2\,\mathbb{P}(\nu)}. \tag{A.62}$$

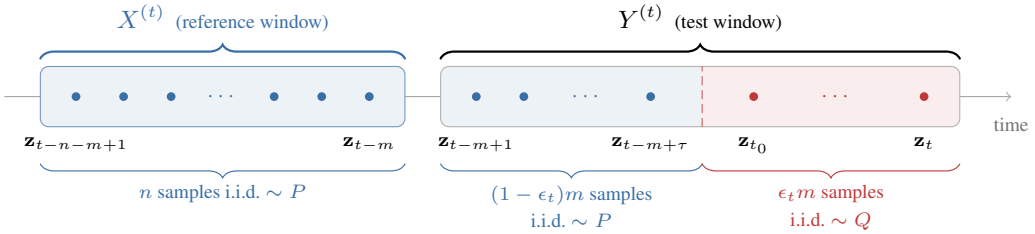


Figure 4: Illustration of the consecutive windows in the contaminated regime $t_0 \le t < t_0 + m$: the samples in the reference window $X^{(t)}$ are drawn i.i.d. from P, while the test window $Y^{(t)}$ contains $\tau = (1-\epsilon_t)m$ samples drawn i.i.d. from P followed by $\epsilon_t m$ samples drawn i.i.d. from Q, where $\epsilon_t = (t - t_0 + 1)/m \in (0, 1]$ is the fraction of post-change samples. The change point $t_0 = t - m + \tau + 1$ is the index of the first post-change sample in the test window.

Combining with (A.59) yields the conservative ARL lower bound

$$\mathbb{E}_\infty[T] \;\ge\; \frac{1}{6}\exp\left(\left(\frac{\nu - C_{n,m}C_d}{3\sqrt{2}\,C_{n,m}}\right)^2\right), \tag{A.63}$$

which is valid whenever $\mathbb{P}(\nu) \le 1$, i.e., for $\nu \ge C_{n,m}\left(C_d + 3\sqrt{2}\sqrt{\ln 3}\right)$.

From (A.63), it suffices to enforce $\frac{1}{6}\exp(((\nu - C_{n,m}C_d)/(3\sqrt{2}C_{n,m}))^2) \ge \gamma$, which is equivalent to the threshold condition $\nu \ge C_{n,m}\big(C_d + 3\sqrt{2}\,\sqrt{\ln(6\gamma)}\big)$ of Theorem 4.9. The auxiliary condition $\nu \ge C_{n,m}\left(C_d + 3\sqrt{2}\sqrt{\ln 3}\right)$ ensuring $\mathbb{P}(\nu) \le 1$ is implied by $\gamma \ge \frac{1}{2}$. □

Theorem 4.10 (False alarm probability). *Let $\alpha \in (0, 1)$, and let T be the stopping time* (23) *with the time-varying thresholds $\nu_t := C_{n,m}\big(C_d + 3\sqrt{2}\,\sqrt{\ln(3t(t+1)/\alpha)}\big)$. Under Assumptions 4.1–4.3, T satisfies the infinite-horizon false-alarm guarantee $\mathbb{P}_\infty(T < \infty) \le \alpha$.*

Proof. Recall the alarm events $A_t := \{g_t > \nu_t\}$ and note that $\{T < \infty\} = \bigcup_{t=1}^{\infty} A_t$ by the definition (23) of T. By the union bound without assuming independence across tests,

$$\mathbb{P}_\infty(T < \infty) = \mathbb{P}_\infty\left(\bigcup_{t=1}^{\infty} A_t\right) \le \sum_{t=1}^{\infty} \mathbb{P}_\infty(A_t). \tag{A.64}$$

It therefore suffices to allocate a per-time significance budget α_t such that $\sum_{t=1}^{\infty} \alpha_t \le \alpha$ and to choose the thresholds ν_t so that $\mathbb{P}_\infty(A_t) \le \alpha_t$ for every t. We take

$$\alpha_t = \frac{\alpha}{t(t+1)}, \tag{A.65}$$

for which $\sum_{t=1}^{\infty} \alpha_t = \alpha$ since $\sum_{t=1}^{\infty} \frac{1}{t(t+1)} = 1$. The tail bound (A.58) is available at every test index t, and the thresholds ν_t of Theorem 4.10 satisfy $\nu_t \ge C_{n,m}C_d$ for all t, hence

$$\mathbb{P}_\infty(A_t) = \mathbb{P}_\infty(g_t > \nu_t) \le 3\exp\left(-\left(\frac{\nu_t - C_{n,m}C_d}{3\sqrt{2}\,C_{n,m}}\right)^2\right). \tag{A.66}$$

With the choice of ν_t in Theorem 4.10, the right-hand side equals

$$3\exp\left(-\ln\left(\frac{3t(t+1)}{\alpha}\right)\right) = \frac{\alpha}{t(t+1)} = \alpha_t. \tag{A.67}$$

Therefore $\mathbb{P}_\infty(A_t) \le \alpha_t$ for all t, and combining with (A.64) and $\sum_{t=1}^{\infty} \alpha_t = \alpha$ yields $\mathbb{P}_\infty(T < \infty) \le \alpha$. □

Lemma A.4. *Let t be a window position in the contaminated regime $t_0 \le t < t_0 + m$, let $\epsilon_t := (t - t_0 + 1)/m \in (0, 1]$ be the fraction of post-change samples in the right window $Y^{(t)}$, and define the margin $\Delta_t := \epsilon_t W_{\mathrm{NN}}(P, Q) - C_{n,m}C_d - \nu_t$. Under Assumptions 4.1–4.3, the miss event $\{g_t \le \nu_t\}$ satisfies*

$$\mathbb{P}(g_t \le \nu_t) \;\le\; \exp\Big(- \frac{2(\Delta_t)_+^2}{c_{xy}^2}\Big), \tag{A.68}$$

where $(u)_+ := \max\{u, 0\}$.

Proof. Fix a window position t in the contaminated regime $t_0 \le t < t_0 + m$, and let $\tau := t_0 - t + m - 1 \in \{0, \dots, m-1\}$ denote the number of pre-change samples in the right window $Y^{(t)} = \{\mathbf{z}_{t-m+1}, \dots, \mathbf{z}_t\}$ (see Figure 4), so that the first post-change sample in $Y^{(t)}$ is $\mathbf{z}_{t-m+\tau+1} = \mathbf{z}_{t_0}$, i.e.,

$$t_0 = t - m + \tau + 1. \tag{A.69}$$

Accordingly, $Y^{(t)}$ consists of τ samples drawn i.i.d. from P followed by $m - \tau$ samples drawn i.i.d. from Q, and the fraction of post-change samples in $Y^{(t)}$ is

$$\epsilon_t = \frac{t - t_0 + 1}{m} = \frac{m - \tau}{m} \in (0, 1], \tag{A.70}$$

so that $(1 - \epsilon_t)m = \tau$ is an integer. The left window $X^{(t)} = \{\mathbf{z}_{t-n-m+1}, \dots, \mathbf{z}_{t-m}\}$ ends at index $t - m < t_0$ and hence contains only pre-change samples. This is exactly the setting of Theorem 4.7 with contamination degree $\epsilon = \epsilon_t$, so Corollary 4.8 applied to $g_t = \widehat{W}_{\mathrm{NN}}[X^{(t)} \,|\, Y^{(t)}]$ gives, for any $\delta \in (0, 1)$, with probability at least $1 - \delta$,

$$g_t \ge \epsilon_t W_{\mathrm{NN}}(P, Q) - C_{n,m}C_d - \sqrt{2}\, B \sqrt{\left(\frac{1}{n} + \frac{1}{m}\right) \ln(1/\delta)}\,, \tag{A.71}$$

where we used that $C_{n,m}C_d = 2B \prod_{\ell=1}^{d} M_\ell \big(\sqrt{2\log(2)\, d} + 1\big)\big(\frac{1}{\sqrt{n}} + \frac{1}{\sqrt{m}}\big)$ is exactly the Rademacher term of Corollary 4.8. Subtracting ν_t from both sides and using the definition of the margin Δ_t in Lemma A.4 gives

$$g_t - \nu_t \;\ge\; \Delta_t - \sqrt{2}\, B \sqrt{\left(\frac{1}{n} + \frac{1}{m}\right) \ln(1/\delta)}\,. \tag{A.72}$$

Hence, on the event $\{g_t \le \nu_t\}$ we must have

$$\sqrt{2}\, B \sqrt{\left(\frac{1}{n} + \frac{1}{m}\right) \ln(1/\delta)} \;\ge\; \Delta_t. \tag{A.73}$$

If $\Delta_t > 0$, choose $\delta := \exp\big(- 2\Delta_t^2 / c_{xy}^2\big) \in (0, 1)$ with $c_{xy}^2 = 4B^2\big(\frac{1}{n} + \frac{1}{m}\big)$, which makes the right-hand side equal to Δ_t and thus implies $\mathbb{P}(g_t \le \nu_t) \le \delta = \exp\big(- 2(\Delta_t)_+^2 / c_{xy}^2\big)$. If $\Delta_t \le 0$, then $(\Delta_t)_+ = 0$ and the right-hand side of (A.68) equals 1, so the bound holds trivially. □

Theorem 4.11 (Detection delay). *For an integer $1 \le k \le m$, define the margin $\Delta_{t_k} := \epsilon_{t_k} W_{\mathrm{NN}}(P, Q) - C_{n,m}C_d - \nu_{t_k}$, and let c_{xy} be as in Theorem 4.7. Under Assumptions 4.1–4.3,*

$$\mathbb{P}(D \ge k) \le \exp\Big(- \frac{2(\Delta_{t_k})_+^2}{c_{xy}^2}\Big), \tag{24}$$

where $(u)_+ := \max\{u, 0\}$.

Proof. Recall that $t_s = t_0 + s - 1$ denotes the s-th window position after the change begins affecting the right window, and let $A_s := \{g_{t_s} > \nu_{t_s}\}$ be the event of an alarm at t_s, so that $D = \inf\{s \ge 1 :\ A_s \text{ occurs}\} - 1$ and $\{D \ge k\} = \bigcap_{s=1}^{k} A_s^c$. We do *not* assume the events $\{A_s^c\}_{s \ge 1}$ are independent (they are typically dependent due to overlapping windows), hence in general one

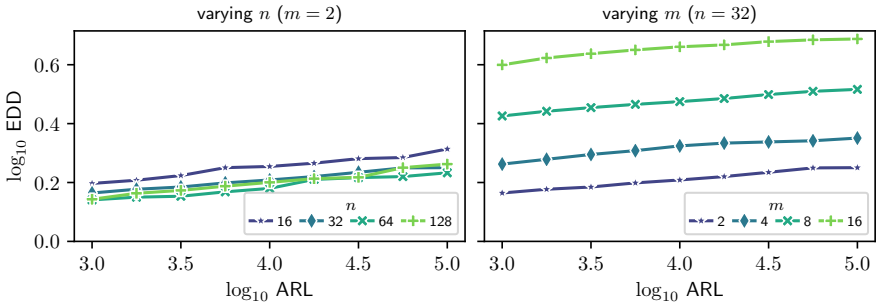


Figure 5: Ablation study on the window sizes of WONDER on the synthetic data.

cannot write $\mathbb{P}(\cap_{s=1}^{k} A_s^c) = \prod_{s=1}^{k} \mathbb{P}(A_s^c)$. Instead, we use the fact that the intersection is contained in each of its constituent events, in particular the last one:

$$\{D \geq k\} = \bigcap_{s=1}^{k} A_s^c \subseteq A_k^c, \qquad \text{hence} \qquad \mathbb{P}(D \geq k) \leq \mathbb{P}(A_k^c) = \mathbb{P}(g_{t_k} \leq \nu_{t_k}). \tag{A.74}$$

We choose the last event A_k^c because the contamination fraction $\epsilon_{t_s} = s/m$ increases linearly in s while the threshold ν_{t_s} of Theorem 4.10 grows only logarithmically in t_s, so the margin Δ_{t_s} is increasing in s in all practical regimes and $s = k$ yields the tightest bound. It remains to apply Lemma A.4 at $t_k = t_0 + k - 1$. Since $1 \leq k \leq m$, the right window $Y^{(t_k)} = \{\mathbf{z}_{t_k-m+1}, \ldots, \mathbf{z}_{t_k}\}$ contains exactly the k post-change samples $\mathbf{z}_{t_0}, \ldots, \mathbf{z}_{t_k}$ preceded by $m - k$ pre-change samples, so by (A.70) the contamination fraction is $\epsilon_{t_k} = (t_k - t_0 + 1)/m = k/m$; moreover, the left window $X^{(t_k)}$ ends at $\mathbf{z}_{t_k-m} = \mathbf{z}_{t_0+k-1-m}$ with $t_0 + k - 1 - m < t_0$, so it contains only pre-change samples and t_k lies in the contaminated regime. Lemma A.4 therefore yields $\mathbb{P}(g_{t_k} \leq \nu_{t_k}) \leq \exp\big(-2(\Delta_{t_k})_+^2/c_{xy}^2\big)$, and (24) follows. □

A.3 Additional Experiments

A.3.1 Ablation Study

We study the sensitivity of WONDER to the size n of the reference window and the size m of the test window on the three synthetic scenarios of Section 5.1, under the same evaluation protocol and with all other hyperparameters unchanged. We vary $n \in \{16, 32, 64, 128\}$ with $m = 2$ fixed, and $m \in \{2, 4, 8, 16\}$ with $n = 32$ fixed; the thresholds of each configuration are recalibrated by Monte Carlo simulation. Figure 5 reports the EDD averaged over the three post-change distributions as a function of the target ARL.

The detection delay is essentially insensitive to the reference window size once $n \geq 32$: the curves for $n = 32$, 64, and 128 lie within 0.1 of each other in $\log_{10}$ EDD over the whole ARL range, whereas the smallest window $n = 16$ is noticeably slower, in particular at large target ARL. This motivates the choice $n = 32$ in Section 5.1, the smallest window that incurs no loss in delay, which also reduces the cost of each gradient step.

The test window size has a much stronger effect: the EDD increases steadily with m at every target ARL, and $m = 2$ yields the shortest delay. This is consistent with the analysis of Section 4.3: in the contaminated regime, the statistic separates from its null level in proportion to the contamination fraction $\epsilon_t = (t - t_0 + 1)/m$ in (A.70), so a larger test window requires more post-change samples before the statistic exceeds the threshold. For all configurations, the EDD grows only slowly with the target ARL, as observed in the main experiments.

A.3.2 MNIST data with post-change digits 4–9

Figure 6 complements Figure 3 by reporting the average detection delay as a function of the target ARL for the MNIST experiments with post-change digits 4–9, under the same experimental protocol as in Section 5.2. The conclusions are unchanged: WONDER attains the lowest EDD on every digit

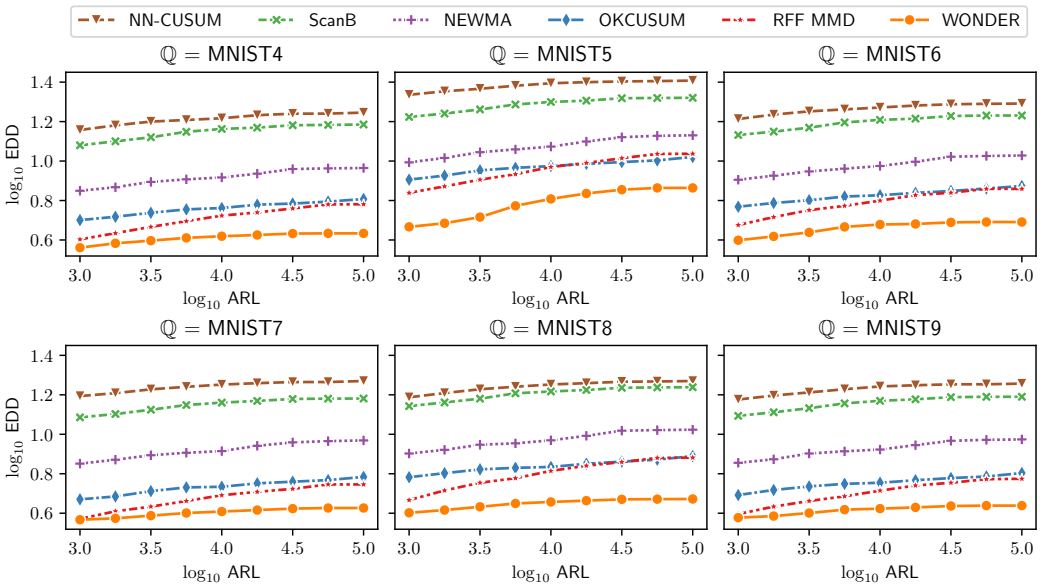


Figure 6: MNIST data: average detection delay versus target ARL from digit 0 to digits 4–9, averaged over 100 runs.

across the considered ARL range, with nearly flat EDD curves, and the largest margins on the harder digits (e.g., digits 5, 6, and 8).

A.4 External statements

We gather here some existing results used in the proofs to ensure self-completeness. Theorem A.5 recalls McDiarmid's inequality from (Boucheron et al., 2013, Section 6.1), which is also known as the bounded differences inequality (McDiarmid, 1989).

Theorem A.5. *Let $\mathcal{X}$ be a measurable space. A function $f : \mathcal{X}^n \to \mathbb{R}$ has the bounded difference property for some constants $c_1, \ldots, c_n$ if, for each $i = 1, \ldots, n$,*

$$\sup_{\substack{x_1,\ldots,x_n \\ x_i' \in \mathcal{X}}} |f(x_1, \ldots, x_{i-1}, x_i, x_{i+1}, x_n) - f(x_1, \ldots, x_{i-1}, x_i', x_{i+1}, \ldots x_n)| \leq c_i. \tag{A.75}$$

Then, if $X_1, \ldots, X_n$ is a sequence of independently distributed random variables and (A.75) holds, putting $Z = f(X_1, \ldots, X_n)$ and $\nu = \frac{1}{4}\sum_{i=1}^n c_i^2$ for any $t > 0$, it holds that

$$\mathbb{P}(Z - \mathbb{E}(Z) > t) \leq \exp\left(-\frac{2t^2}{\sum_{i=1}^n c_i^2}\right). \tag{A.76}$$